Universitat Autònoma de Barcelona

IMPACT OF TRANSPOSABLE ELEMENTS IN THE EVOLUTION OF PLANT GENOMES

Cristina Vives Cobo

Universitat Autònoma de Barcelona

Facultat de Biociències

Departament de Biologia Animal, Biologia Vegetal i Ecologia

Programa de Doctorat en Biologia i Biotecnologia Vegetal

TESIS DOCTORAL

IMPACT OF TRANSPOSABLE ELEMENTS IN THE EVOLUTION OF PLANT GENOMES

Memòria presentada per Cristina Vives i Cobo per optar al títol de doctora per la Universitat Autònoma de Barcelona.

El treball s'ha realitzat en el Programa de Genòmica de Plantes i Animals del Centre de Recerca en Agrigenòmica (CRAG), Campus UAB Bellaterra, sota la direcció del doctor Josep M. Casacuberta Suñer.

El director de tesis:	La tutora de tesis:
Dr. Josep M. Casacuberta Suñer	Dra. Roser Tolrà Pérez

La candidata a doctora:

Cristina Vives Cobo

Barcelona, 2017

Als meus pares

A la Núria

A l'Ester

INDEX

List of tables

List of figures

Abbreviations

AP	Aspartic Protease
BLAST	Basic Local Alignment Search Tool
BLR	BLASTER
BS	Binding Site
Bur	Burren
ChIP	Chromatin ImmunoPrecipitation analysis
Chr	Chromosome
CL	‘Chinese Long’ melon accession
Col	Columbia
CV	‘Cabo Verde’ melon accession
DEL	Deletion
DNA	Deoxyribonucleic acid
Ds	Dissociator
E2F-TE	E2F binding site within an annotated Transposable Element
GO	Gene Ontology
InDel	Insertion-Deletion polymorphism
INS	Insertion
INT	Integrase
Kro	Krotzenburg
LARD	LArge Retrotransposons Derivative
Ler	Landsberg erecta
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MITE	Miniature Inverted-repeat Transposable Element
Mya	Million years ago
PAV	Presence-Absence Variation
PCR	Polymerase Chain Reaction
PM-TE	TE-related Polymorphic loci
PS	‘Piel de sapo’ melon accession
RNA	Ribonucleic acid
RT	Reverse Transcriptase

RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SC	‘Songwhan Charmi’ melon accession
SINE	Short Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SSAP	Sequence-Specific Amplification Polymorphism
SSR	Simple Sequence Repeat
SV	Structural Variation
TE	Transposable Element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TIR	Terminal Inverted Repeat
TPase	Transposase
TSD	Target Site Duplication
WT	Wild Type
WU-BLAST	Washington University - BLAST

Units

bp	Base pairs
Gb	Gigabase
Kb	Kilobase
Mb	Megabase
Nt	nucleotides

SUMMARY

Transposable elements are genetic elements that have the capacity to modify their position within the genome. As a consequence, they impact the evolution of genomes by inactivating or altering host genes and by providing new gene functions. Transposons account for an important fraction of all sequenced genomes. The goal of the work presented in this dissertation is to investigate the diverse impacts of transposons on gene and genome evolution in different plant species.

The transposon content has been analyzed in melon and cucumber, two closely related species. The results suggest that transposons have proliferated to a greater extend in melon, causing an increase of its genome size. Transposable elements are usually not homogenously distributed and tend to accumulate in heterochromatic pericentromeric regions. This is also the case of melon and cucumber genomes. Interestingly, the results presented show that transposons have expanded the pericentromeric regions in melon, showing that transposons can modify the structure of genomes.

The number of plant reference genomes made available and the number of varieties resequenced is growing exponentially, and this is allowing to study the correlation between genetic and phenotypic variations. The purpose of the work summarized in the second part of this dissertation is to analyze the impact of transposons in crop genomes by detecting polymorphisms due to the presence or absence of transposon at a given locus, comparing one resequenced variety respect to the reference genome. The analysis of transposon-related polymorphism insertions has been performed in three different species: melon, date palm and *Physcomitrella patens*. The results obtained can help to identify the transposon families recently active and to provide new information on genetic polymorphisms that can be linked to traits selected during the recent evolution of these three species.

In order to study the impact of transposition on gene regulation, the work reported in the third part of this dissertation focuses on the capacity of transposons to amplify and redistribute transcription factor binding sites. The results show that some MITE families have amplified and redistributed the binding sites of E2F transcription factor during *Brassica* evolution. The goal of this study was to assess the impact of E2F binding sites located within a transposon in reprogramming gene regulation on the E2F

transcriptional network. The results obtained have determined that E2F binding sites located within transposons have the capacity to bind E2F transcription factor *in vivo*, regardless the epigenetic mark context.

Moreover, transposons have become a useful genetic tool to generate mutant collections in animals and plants due to the capacity to insert copies into the genome. In plants, some retrotransposons have been shown to integrate preferentially near genes making them particularly interesting for mutagenesis. Among them, the tobacco retrotransposon Tnt1 has been used to generate mutants in different plant species.

The last part of this dissertation consists in analyzing the capacity of the tobacco retrotransposon Tnt1 to transpose in the moss *Physcomitrella patens*. It shows that Tnt1 efficiently transposes in *P. patens* and inserts preferentially in genic regions. This work presents Tnt1-derived vectors designed for high efficiency transposition that could be used to generate stable insertion mutant collections in this bryophyte species.

RESUM

Els transposons són elements genètics que tenen la capacitat de modificar la seva posició dins el genoma. Com a conseqüència, tenen un impacte en l'evolució del genomes inactivant o alterant els gens de l'hoste i proporcionant noves funcions gèniques. Els transposons ocupen una fracció important de tots els genomes seqüenciats. L'objectiu del treball presentat en aquesta tesis consisteix en estudiar els diversos impactes de transposons tant en els gens com en l'evolució dels genomes de diferents espècies de plantes.

En aquesta tesis, s'ha analitzat la fracció de transposons en meló i cogombre, dues espècies molt properes. Els resultats suggereixen que els transposons han proliferat més en meló, causant un augment de la mida del genoma. Els transposons no es troben distribuïts habitualment de forma homogènia i tendeixen a acumular-se en les regions pericentromèriques heterocromàtiques, com el cas dels genomes de meló i cogombre. Curiosament, els resultats presentats mostren que els transposons han expandit les regions pericentromèriques en meló, demostrant que els transposons poden modificar l'estructura dels genomes.

El número de genomes de referència de plantes disponibles i el número de varietats reseqüenciades ha crescut exponencialment permetent estudiar la correlació entre les variacions genètiques i fenotípiques. El propòsit del treball resumit en la segona part d'aquesta tesis consisteix en analitzar l'impacte dels transposons en genomes d'espècies cultivables detectant els polimorfismes deguts a la presència o absència de transposó en un locus concret, comparant una varietat reseqüenciada respecte al seu genoma de referència. L'anàlisi d'insercions polimòrfiques de transposons s'ha realitzat en tres espècies diferents: meló, palmera datilera i *Physcomitrella patens*. Els resultats obtinguts poden ajudar a identificar famílies de transposons actives recentment i proporcionar informació nova sobre polimorfismes genètics que poden estar lligats a caràcters seleccionats durant l'evolució recent d'aquestes tres espècies.

Per tal d'estudiar l'impacte de la transposició en la regulació gènica, el treball presentat en la tercera part d'aquesta tesis se centra en la capacitat dels transposons en amplificar i redistribuir llocs d'unió a factors de transcripció. Els resultats mostren que algunes famílies de MITEs s'han amplificat i han redistribuït els llocs d'unió del factor de

transcripció E2F durant l'evolució d'algunes espècies del gènere *Brassica*. L'objectiu d'aquest treball és avaluar l'impacte dels llocs d'unió a E2F localitzats dins de transposons reprogramant la regulació de gens de la xarxa transcripcional de E2F. Els resultats obtinguts han determinat que els llocs d'unió a E2F localitzats dins de transposons tenen la capacitat d'unir-se als factors de transcripció de E2F *in vivo*, independentment de les marques epigenètiques de la regió.

A més a més, els transposons s'han convertit en eines genètiques útils per generar col·leccions de mutants en animals i plantes degut a la seva capacitat d'integrar còpies en el genoma. En plantes, alguns retrotransposons s'integren preferentment a prop de gens sent particularment interessants per la mutagènesis. Entre tots ells, el retrotransposó de tabac Tnt1 s'ha utilitzat per generar mutants en diferents espècies de plantes.

L'última part d'aquesta tesis consisteix en analitzar la capacitat del retrotransposó de tabac Tnt1 en transposar en la molsa *Physcomitrella patens*. S'ha demostrat que Tnt1 transposa eficientment en *P. patens* i s'integra preferentment a prop de gens. Aquest estudi presenta vectors derivats de Tnt1 dissenyats per transposar amb alta eficiència i ser utilitzats per generar col·leccions de mutants amb insercions estables en aquest briòfit.

RESUMEN

Los transposones son elementos genéticos que tienen la capacidad de modificar su posición en el genoma. Como consecuencia, tienen un impacto en la evolución de los genomas inactivando o alterando los genes del huésped y proporcionando nuevas funciones génicas. Los transposones ocupan una fracción importante en todos los genomas resecuenciados. El objetivo del trabajo presentado en esta tesis trata en estudiar los distintos impactos de transposones tanto en genes como en la evolución de los genomas de distintas especies de plantas.

En esta tesis, se ha analizado la fracción de transposones en melón y pepino, dos especies muy cercanas. Los resultados sugieren que los transposones han proliferado más en melón, causando un aumento del tamaño del genoma. Los transposones no se encuentran distribuidos habitualmente de forma homogénea y tienden a acumularse en las regiones pericentroméricas heterocromáticas, como es el caso de los genomas de melón y pepino. Curiosamente, los resultados presentados muestran que los transposones han expandido las regiones pericentroméricas en melón, demostrando que los transposones pueden modificar la estructura de los genomas.

El número de genomas de referencia de plantas disponibles y el número de variedades resecuenciadas ha crecido exponencialmente permitiendo estudiar la correlación entre las variaciones genéticas y fenotípicas. El propósito del trabajo resumido en la segunda parte de esta tesis consiste en analizar el impacto de los transposones en genomas de especies cultivables, detectando los polimorfismos ocasionados por la presencia o ausencia de transposones en un locus concreto, a través de la comparación de una variedad resecuenciada respecto al genoma de referencia. El análisis de inserciones polimórficas de transposones se ha realizado en tres especies distintas: melón, palmera datilera y *Physcomitrella patens*. Los resultados obtenidos pueden ayudar a identificar familias de transposones activas recientemente y proporcionar información nueva sobre polimorfismos genéticos que pueden estar ligados a caracteres seleccionados durante la evolución reciente de estas tres especies.

Para estudiar el impacto de la transposición en la regulación génica, el trabajo presentado en la tercera parte de esta tesis se centra en la capacidad de los transposones en amplificar y redistribuir sitios de unión a factores de transcripción. Los resultados muestran que algunas familias de MITEs se han amplificado y redistribuido los sitios de unión del factor de transcripción E2F durante la evolución de algunas especies del género *Brassica*. El objetivo de este trabajo ha sido evaluar el impacto de los sitios de unión a E2F localizados dentro de transposones reprogramando la regulación de genes en la red transcripcional de E2F. Los resultados obtenidos han determinado que los sitios de unión a E2F localizados dentro de transposones tienen capacidad de unirse a los factores de transcripción de E2F *in vivo*, independientemente de las marcas epigenéticas en la región.

Además, los transposones se utilizan como herramienta genética útil para generar colecciones de mutantes en animales y plantas debido a su capacidad de integrar copias en el genoma. En plantas, algunos retrotransposones se integran preferentemente cerca de genes siendo particularmente interesantes para la mutagénesis. De entre todos, el retrotransposón de tabaco Tnt1 se ha utilizado para generar mutantes en distintas especies de plantas.

La última parte de esta tesis consiste en analizar la capacidad del retrotransposón de tabaco Tnt1 en transponer en el musgo *Physcomitrella patens*, ya que se demostró que Tnt1 transpone eficientemente en *P. patens* y se integra preferentemente cerca de genes. Finalmente, este estudio presenta vectores derivados de Tnt1 diseñados para transponer con alta eficiencia y ser utilizados para generar colecciones de mutantes con inserciones estables en esta especie briofita.

GENERAL INTRODUCTION

GENERAL INTRODUCTION

Transposable elements: definition and classification

Transposable elements (TEs) are genetic elements that have the capacity to modify their position within the genome and, in some cases, to generate new copies of themselves. As a consequence, TEs are an important source of mutations and they account for an important fraction of all sequenced genomes (Tenaillon et al. 2010).

In the 1940s, Barbara McClintock was the first person to postulate the existence of "controlling elements" which could alter gene expression by their movement. Her work consisted in studying the relationship between chromosome breaks and grain color variability in maize (McClintock 1947 and 1948). She established that the chromosome breaks were linked to the presence of a factor named *Dissociator* (*Ds*) and she demonstrated that the element *Ds* could only generate the breaks in the presence of another element named *Activator* (*Ac*) (McClintock 1953). This system involving these two elements of control was named system *Ac*/*Ds*. But, it was not until the 1980s that *Ac* and *Ds* elements were molecularly cloned and characterized (Fedoroff 1989). Since then, many TEs have been identified and characterized in different organisms.

TEs are a very diverse group of genetic elements and can be classified based on their structure and mode of transposition. The most widely used classification is the one proposed by Wicker et al. (2007). At the highest level, TEs can be classified into two major classes, class I (retrotransposons) and class II (DNA transposons). Within each class, TEs are further subdivided in orders, depending on their insertion mechanism, encoded proteins and structure, in superfamilies, based on their replication strategy and in families, based on sequence similarity (Wicker et al. 2007; Kapitonov and Jurka 2008). For both class I and class II, genomes contain autonomous elements, which encode for the proteins needed for their transposition, and non-autonomous elements, which contain the *cis*-elements required for transposition but lack some of the proteins needed and can only transpose using the proteins provided in *trans* by other elements (Figure 1).

Class I elements, or retrotransposons, transpose via an RNA intermediate, through a 'copy-and-paste' mechanism, leaving a copy behind and integrating a new copy in a different genomic location (Figure 2a). Elements within this class are subdivided into elements with Long Terminal Repeats (LTRs), known as LTR retrotransposons, and without LTRs, known as non-LTR retrotransposons.

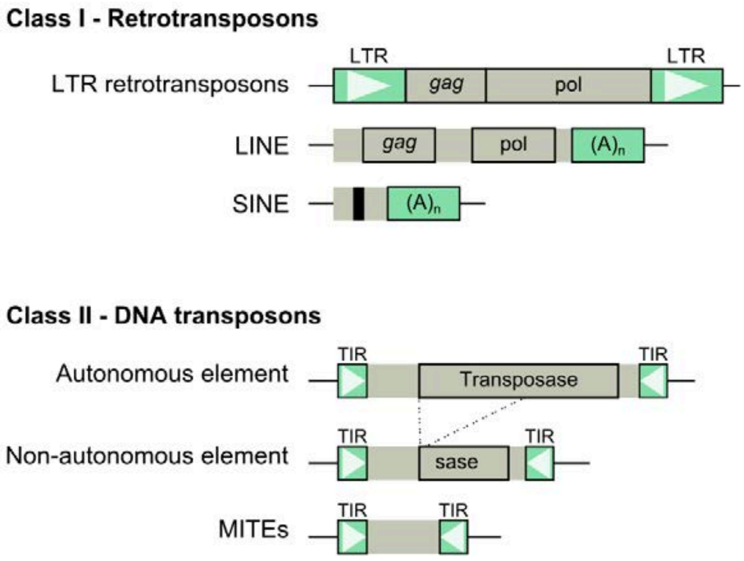


Figure 1. Schematic structure of the different types of plant transposable elements. The black box in SINEs stands for the tRNA-related region. Adapted from Casacuberta and Santiago 2003.

Together with MITEs (see below), LTR retrotransposons are the most common TEs in plants (Kumar and Bennetzen 1999; Casacuberta and Santiago 2003). The transcription of these elements starts in the 5' LTR and ends at the 3' LTR. The LTRs usually contain the promoter and transcriptional regulatory elements. Autonomous LTR retrotransposons contain two major genes: *gag* and *pol*. *Gag* proteins are structural proteins essential to form the virus-like particle, whereas the *pol* gene encodes the proteins needed for the retrotransposon life cycle. The first step of retrotransposition is transcription. The RNA transcript of an integrated copy is used as a template to make a new DNA copy by RNAseH and reverse transcriptase (RT) and then the integrase (INT) allows the insertion of the double-stranded DNA back into the host genome (Figure 2a). There are two main superfamilies of LTR retrotransposons, *Gypsy* and *Copia*, which

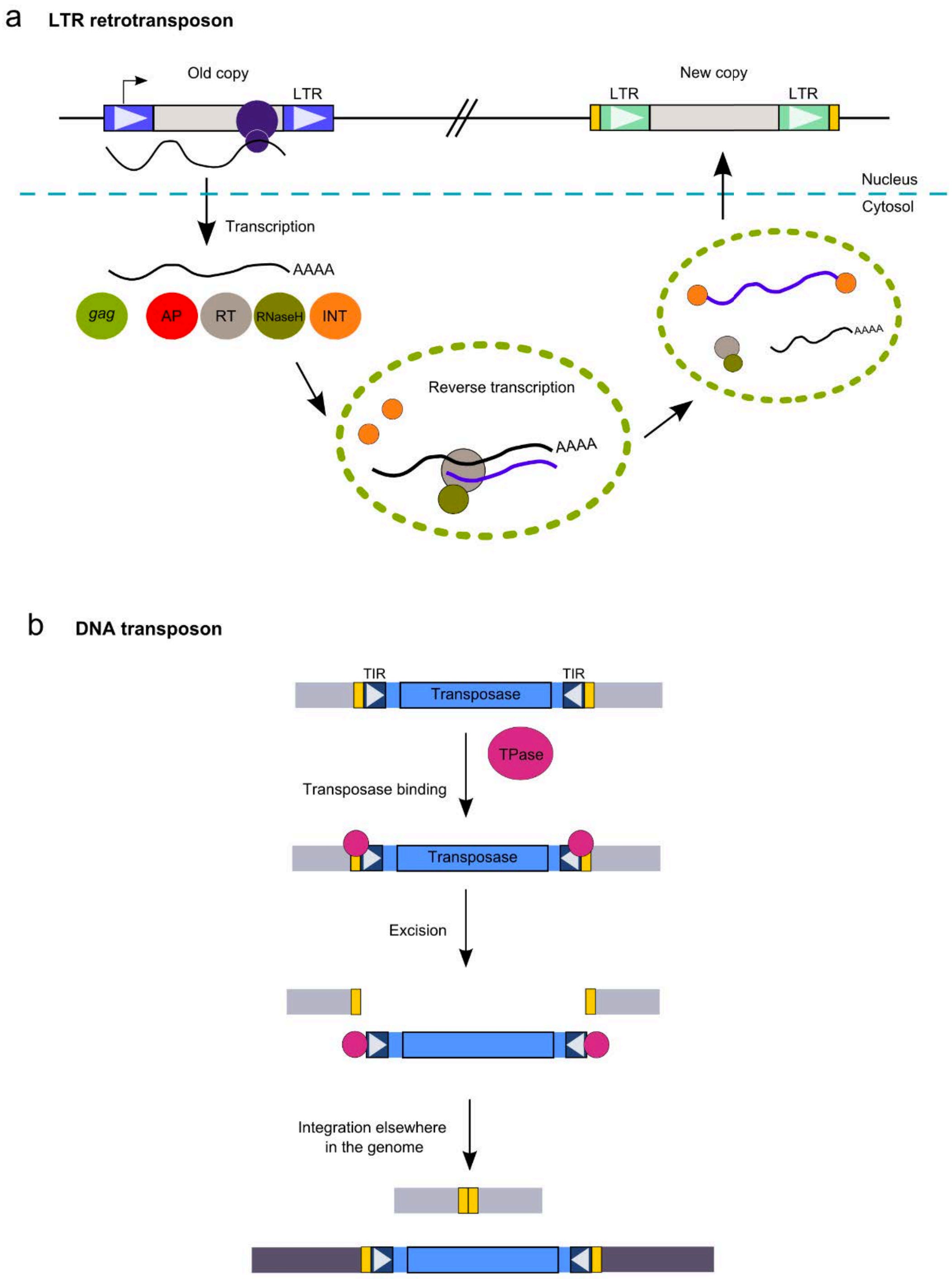


Figure 2. Mechanisms of transposition for the two main classes of mobile elements. (a) Life cycle of a LTR retrotransposon. The order of the polyprotein corresponds to a *Gypsy* element, and encodes for *Gag*, aspartic proteinase (AP), reverse transcriptase (RT), RNAseH and integrase (INT). (b) Mobilization of a DNA transposon. The yellow boxes represent TSDs.

differ in the order of protein domains, as well as in some other characteristics. Whereas INT precedes RT and RNAseH in *Copia* superfamily elements, INT is the last one in *Gypsy* superfamily elements.

Some non-autonomous elements derived from LTR retrotransposons are LArge Retrotransposons Derivatives (LARDs) with large internal sequence region between the two LTRs (Kalendar et al. 2004), whereas others are qualified as Terminal-repeat Retrotransposons In Miniature (TRIMs) with a few hundred bps of internal sequence (Witte et al. 2001; Gao et al. 2012).

The non-LTR retroelements present a repetitive sequence at the end of 3', instead of being flanked by LTRs. Depending on their coding capacity, they can be further subdivided into Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). The most common non-LTR retroelements in plants are LINEs (Lisch 2013), which usually code for *gag* protein and a *pol* polyprotein, including an endonuclease and a RT. In contrast, SINEs are defective elements with no coding capacity and require the proteins from autonomous LINEs to transpose. Neither LINEs nor SINEs are frequent in plant genomes, but they have proliferated in some mammalian genomes.

Class II elements, or DNA transposons, transpose by excising from their position and integrating at a different genomic location, by a mechanism known as 'cut-and-paste' (Figure 2b). These elements can be flanked by short Terminal Inverted Repeats (TIRs). The autonomous elements encode for a transposase which allows the excision, frequently recognizing the TIRs, whereas defective TE copies are mobilized by the transposases encoded by an intact related element. The integration of an element in a new location generates a characteristic Target Site Duplication (TSD).

The classification of DNA transposons is based on the transposase motifs, the TIR sequences and the size and sequence of the TSD. The main superfamilies of DNA transposons are PIF/Harbinger, hAT, Tc1/Mariner, CACTA, MULE. Apart from these characteristics, some DNA transposons, named *Helitrons*, can be mobilized via a 'rolling-circle' mechanism similar to some bacterial TEs. These elements are classified as DNA transposon because they don't transpose through an RNA intermediate.

Moreover, an interesting type of defective DNA transposons are Miniature Inverted-repeat Transposable Elements (MITEs). These elements are very short, between 100 and 700 bp, and frequently contain TIRs of their corresponding autonomous elements which are flanking non-coding internal sequence. In contrast to other defective elements, MITEs can be amplified reaching very high copy numbers (Casacuberta and Santiago 2003; Guermonprez et al. 2012).

TEs are a major component of plant genomes

The TE content differs in plant genomes. For example, 85% of maize genome (Velasco et al. 2010), 82% of barley genome (IBGSC 2012) or 81% of sunflower genome (Natali et al. 2013) have been annotated as TEs, whereas TEs represent a 21% in the compact genome of *A. thaliana* (Ahmed et al. 2011). Although these numbers are not directly comparable, as the annotation methods differ, there is a relationship between genome size and its TE fraction.

Plant genomes contain all major types and classes of TEs, where LTR retrotransposons and MITEs are the most abundant ones (Casacuberta and Santiago 2003). As MITEs are small elements, LTR retrotransposons are mainly responsible for the differences in genome sizes (Bennetzen et al. 2005). For example, they account for only the 2.5% of the small genome (77 Mb, 1C = 0.092 pg) of *U. gibba* (Ibarra-Laclette et al. 2013) and as much as the 90% of the big (1C = 50.9 pg) *Fritillaria* species genome (Ambrozová et al. 2010).

Therefore, TEs, and in particular LTR retrotransposons, have had an enormous impact on the evolution of plant genomes, and in particular on their size. Moreover, transposition activity is not constant overtime. There are some periods where TEs are relatively quiescent and others where transposition bursts increase significantly the copy number (Vitte et al. 2014). These bursts of transposition can affect several or only specific families. This explains that the prevalence of a particular family varies among species and even among varieties of the same species (Kumar 2015; Ambrozová et al. 2010). These bursts of transposition can lead to important changes in genome size in very short time frames. For instance, *Oryza australiensis*, a wild relative of rice, has double its genome size by a transposition burst affecting only few retrotransposons families during the last three million years (Piegu et al. 2006; Zhao and Ma 2013).

TEs affect genome structure

TEs are not homogeneously distributed in most genomes. For example, most plant *Copia*-like TEs tend to insert into euchromatic regions, while most plant *Gypsy*-like TEs show some preference for pericentromeric regions (Peterson-Burch et al. 2004). Besides preferential insertion for some TEs, they tend to be accumulated in pericentromeric regions, which can be explained mainly by two reasons. The first one is that the recombination rate varies depending on different chromosomal regions. A lower recombination rate in pericentromeric regions makes the elimination rate of TEs lower in this region. The second reason is the action of selection against TE insertion within genes, due to their higher deleterious effects (Neumann et al. 2011; Bennetzen and Wang 2014). As a consequence, TEs concentrate in gene-poor regions, and in particular in heterochromatic pericentromeric regions. This pattern of accumulation has important consequences for the structure of genomes, and can also impact their evolution.

TEs are the main target of epigenetic silencing, and they are associated to high methylation and heterochromatic histone variants which prevent their mobility (Ito and Kakutani 2014). The result of this concentration of several epigenetic marks is a particular chromatin structure which allows the functionality of centromeres and the heterochromatin compaction (Wong and Choo 2004). Interestingly, some retrotransposons recognize epigenetic marks from heterochromatin and integrate there, reinforcing the maintenance of the epigenetic marks of heretochromatin state (Gao et al. 2008). In plants, some TEs are almost exclusively found in pericentromeric regions, like centromere-specific CRM family in maize (Jin et al. 2004).

Although TEs concentrate in centromeres and pericentromeric regions in all plants, the heterochromatic pericentromeric regions greatly differ in TE concentration and size. Interestingly, some examples have been recently published. The plant model *A. thaliana* has a compact genome with very small pericentromeric TE-rich regions, but the close relative *Arabis alpina* presents a bigger pericentromeric regions and a higher TE content, suggesting that TEs have expanded its pericentromeric regions (Willing et al.

2015). Moreover, it has been shown that the expanded pericentromeric regions in tomato contain recently evolved genes, suggesting that these regions may allow genes to evolve at a different rate (Jouffroy et al. 2016). Chapter 1 presents an analysis of the TE distribution in melon and its close relative cucumber, suggesting that an expansion of the pericentromeric regions has occurred in the former after the split of the two species.

TEs altering genes and as source of new functions

TEs are an important source of mutations and have an important impact on genome evolution. TEs can modify the coding capacity of genes in many ways. They can disrupt genes by altering the reading frame or by introducing a STOP codon. But also, TEs can be incorporated as a new exon, lead to a truncated transcript, introduce new splice sites and create new alternative spliced variants. They can also modify the regulation of genes by providing new promoters or regulatory sequences, by producing antisense transcripts or by modifying the chromatin epigenetic marks (Casacuberta and González 2013).

But apart from altering genes, a TE or a part of it can be the source of new gene functions (Lisch 2013; Oliver et al. 2013). For instance, several transcription factors derive from class II transposases. This is the case, for example, of the FAR1 transcription factor from maize (Lin et al. 2007) or DAYSLEEPER in *A. thaliana* (Bundock and Hooykaas 2005).

Another interesting impact is the TE capacity to capture and to mobilize genes or gene fragments to a new genomic location (Lisch 2013). This seems to be the case, for example, of the *Helitron* elements in maize (Du et al. 2009).

Impact of TEs in gene regulation

Apart from modifying the coding capacity of genes, TEs can alter gene expression by providing their own regulatory elements or by attracting epigenetic silencing machinery (Contreras et al. 2015). Several TEs preferentially transpose into the 5’ region from a gene, which can modify their expression (Liu et al. 2009; Naito et al. 2009). Many plant TEs contain stress-inducible promoters, like the tobacco retrotransposon Tnt1 which is induced by biotic and abiotic stresses (Grandbastien et al. 2005) or the Arabidopsis ONSEN retrotransposon activated by heat (Cavrak et al. 2014). The insertion of stress inducible TEs close to genes may therefore be a mechanism to confer stress-inducibility to new genes and explore new ways to overcome these difficult situations. It is interesting to note that 33% of genes expressed under stress in maize contain a TE in their promoter region, TEs that in most cases also respond to stress (Makarevitch et al. 2015).

Interestingly, some TEs contain transcription factor binding sites (TFBS), and their movement may put new genes under control of existing transcriptional networks (Rebollo et al. 2012). Chapter 3 describes the analysis of MITE families which seem to have amplified and redistributed the binding sites of E2F transcription factor during *Brassica* evolution.

But TEs can also influence the plant stress responses indirectly. It has been recently shown that the epigenetic status of a TE can regulate the stress responses in *Arabidopsis* through the activity of a TE small RNA (McCue et al. 2012). In fact, the most frequent effect of TE insertion within or close to a gene promoter is its inactivation by epigenetic effects. TEs are methylated and are associated to inactive chromatin, and their insertion close to a gene can induce its silencing (Contreras et al. 2015). For example, the presence of a methylated TE in the promoter region of the *CmWIP1* gene determines sex in melon flowers (Martin et al. 2009) (Figure 3a).

TEs dynamics and the evolution of crop plants

As previous explained, TEs can reshape genomes in different ways, from causing chromosome rearrangements to creating new regulatory elements or modifying the existing ones (Piacentini et al. 2014). For these reasons, TEs are a source of genetic variability essential for evolution (Lisch 2013).

Plant domestication and breeding are a particular type of evolution where selection for agronomical interesting traits is the driving force, and TEs have also been an important source of variability selected by humans. In the last few years, a number of TE-associated mutations linked to alleles selected during crop plant domestication and breeding has been described. For example, some alleles have been selected during maize domestication, such as those responsible for changes in flowering time (Salvi et al. 2007), plant architecture (Studer et al. 2011) or photoperiod sensitivity (Yang et al. 2013).

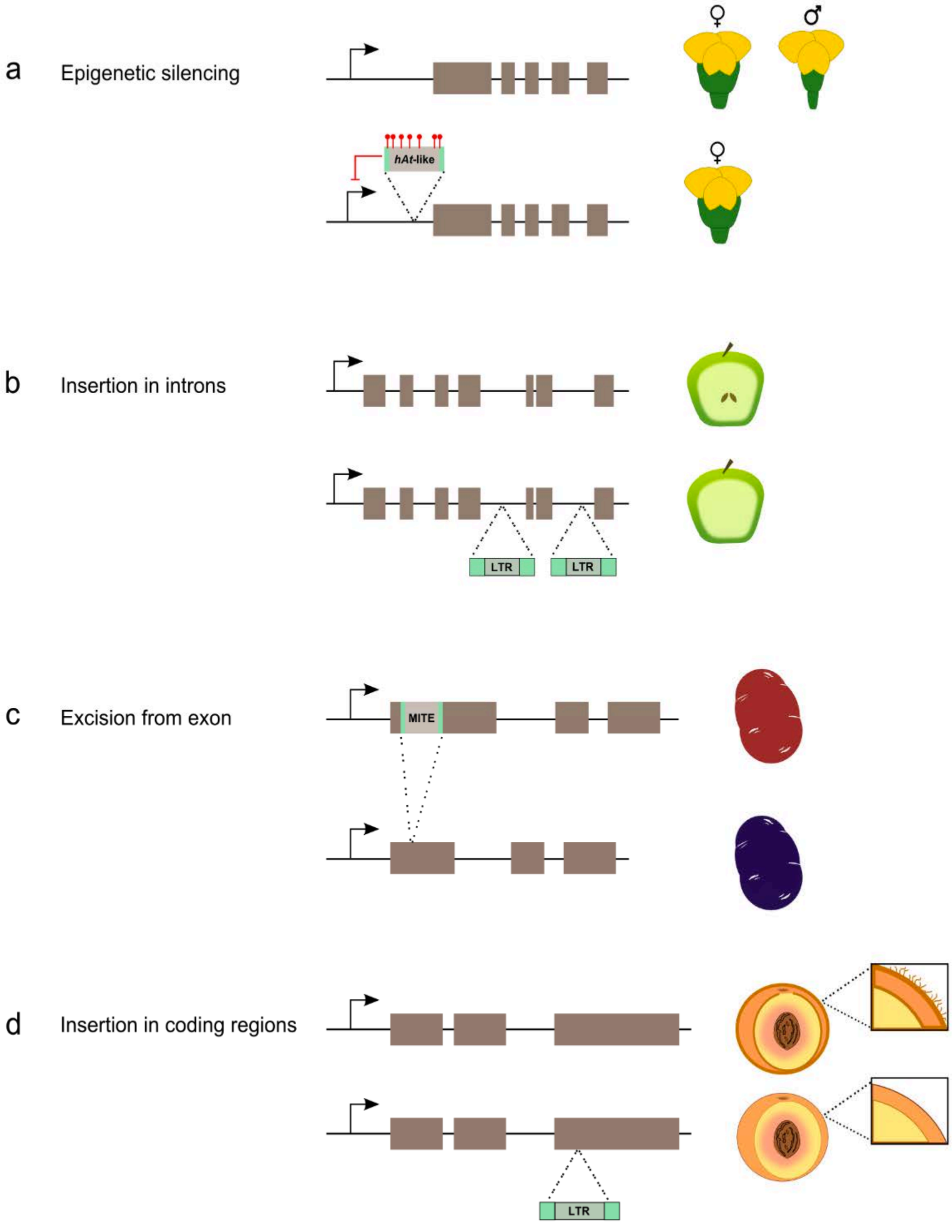


Figure 3. Representation of the functional impact of TE insertions in crops. (a) The sex determination in melon flowers results from epigenetic changes in promoter region of *CmWIP1* gene caused by the insertion of a TE (Martin et al. 2009). (b) LTR retrotransposon insertion in intron 4 or 6 abolishes the expression of the *MdPI* gene and confers the seedless phenotype in some apple varieties (Yao et al. 2011). (c) The excision of MITE from exon allows the expression of flavonoid 3’, 5’ - hydroxylase gene by changing the skin color of potato tubers (Momose et al. 2010). (d) Nectarine phenotype, characterized by the absence of skin pubescence, is due to the insertion of a *Copia* retrotransposon in the third exon of MYB25 gene (Vendramin et al. 2014).

Moreover, some TE insertions have conferred different fruit phenotypes which have been selected during the recent breeding. In grape, a retrotransposon inserted in the promoter region of a *Myb* transcription factor gene confers the loss of pigmentation in the fruit skin, typical for some white grape varieties such as ‘Chardonay’. The recombination between LTRs leaves behind a solo-LTR which has a milder effect on the gene expression conferring an intermediate phenotype with pink grapes (Kobayashi et al. 2004). Another well studied example is the insertion of a *Copia*-like retrotransposon close to the *Ruby* gene in oranges which, after induction by cold stress, regulates anthocyanin production resulting in the blood orange phenotype (Butelli et al. 2012).

Interestingly in peach, the presence of a *Copia* retrotransposon within the third exon of a transcription factor that regulates trichome formation causes the absence of skin pubescence conferring the nectarine phenotype (Vendramin et al. 2014) (Figure 3d). The list of examples of TEs that have impacted crop evolution and breeding increases rapidly (Lisch 2013; Vitte et al. 2014; Contreras et al. 2015) (Figure 3). For instance, it has recently been demonstrated that a MITE inhibits the expression of *Ghd2* gene, which is involved in grain number, plant height and heading date in rice (Shen et al. 2017). And another example, the LTR retrotransposon insertion in BoCYP704B1 gene causes the male sterility in cabbage (Ji et al. 2017).

In the last few years, the genome information on crops and crop varieties has increased exponentially. Since the sequence of the first crop genome, rice, which followed that of the first plant, *Arabidopsis thaliana* in 2000, there are more than 50 complete plant genomes available on the Phytozome database (https://phytozome.jgi.doe.gov/).

In spite of the availability of plant genomes sequenced, the quality of the published genomes is not always good enough for identifying which regions are related to TEs. The variable fraction of the TE content of the sequenced genomes depends on the quality of genome assemblies, and of course the software and parameters used to annotate TEs. Furthermore, according to the objective pursued different TE annotation strategies may be followed: more stringent TE annotation can be useful for evolutionary or phylogenetic studies, while a less-conservative annotation is able to detect more degenerate elements and this approach can be useful for masking genomes or study TE landscape, for example. All these facts make very difficult the comparison of TE content between genomes. For instance, Maumus et al. 2014 combined several repeat annotation programs and increased at least 20% of genome coverage in *A. thaliana* genome compared to annotations obtained from a single program. The challenge is where to put the cutoffs and thresholds and which program to use in order to obtain a TE annotation suitable for each purpose.

In addition to the number of species for which a reference genome is now available and even more importantly, the number of resequenced varieties for a particular crop species has also increased at an exponential rate. As an example, 360 accessions of tomato (Lin et al. 2014) or 3,000 varieties of rice (Li et al. 2014) have been resequenced providing an enormous wealth of information on the mutations linked to important agronomic traits.

However, in most cases the analysis of the variability among species and varieties is limited to SNPs and the variability generated by TE movement is not taken into account. This is mainly due to the intrinsic difficulties to this analysis that requires dedicated bioinformatic tools. Some of these bioinformatic tools have recently been developed in ours and other laboratories, allowing to detect novel TE insertions using paired-end resequencing data. These tools include RetroSeq (Keane et al. 2013), TEMP (Zhuang et al. 2014), T-lex2 (Fiston-Lavier et al. 2015) and Jitterbug (Hénaff et al. 2015). Chapter 2 describes the analysis of TE-related polymorphic loci performed in three different species: melon, date palm and *Physcomitrella patens*. This kind of analyses will allow us to understand the role of TEs generating variability during domestication and breeding processes.

TEs as tools to analyze gene function

Since many plant genomes have been sequenced, reverse genetics can be a suitable strategy to determine the functionality of the large number of new predicted genes. Several reverse genetic approaches have been developed, such as anti-sense or RNAi suppression (Waterhouse and Helliwell 2003), homologous recombination (Schuermann et al. 2005) and insertional mutagenesis (Kumar and Hirochika 2001). Among these, insertional mutagenesis is the most widely used approach for gene function analysis in plants (Ramachandran and Sundaresan 2001). Either T-DNA (Krysan et al. 1999;

Alonso et al. 2003) or transposons (Sundaresan et al. 1995; Walbot 2000) have been used as insertional mutagens in plants. Because of its random integration, T-DNA is not suitable as insertional mutagen in plant species with large genomes (Hou et al. 2010).

However, some TEs, and in particular some retrotransposons, present advantages in specific cases. For instance, the stability of retrotransposon insertions and the preferential insertion close to genes of some retrotransposons make them suitable for mutagenesis in large plant genomes (Kumar and Hirochika 2001). For example, the rice Tos17 retrotransposon has been a great success for the generation of mutant collections in rice (Hirochika 2001; Piffanelli et al. 2007). Apart from using endogenous TEs as insertional mutagens, some TEs are transpositional competence in various heterologous plant, such as the tobacco Tnt1 retrotransposon used in *A. thaliana* (Lucas et al. 1995), *Medicago truncatula* (d'Erfurth et al. 2003) and lettuce (Mazier et al. 2007). Chapter 4 presents the analysis of the capacity of the tobacco Tnt1 retrotransposon to transpose in the bryophyte species *Physcomitrella patens* with the objective to generate stable insertion mutant collections.

CHAPTER 1

Impact of transposons-induced mutations in speciation

CHAPTER 1: Impact of transposons-induced mutations in speciation

Comparison of the transposable elements landscape in two related *Cucumis* species

1.1.- INTRODUCTION

Whole-genome sequencing data has confirmed that TEs account for a quite large fraction of plant genomes (Bennetzen and Wang 2014). TE content is variable depending on plant species, from 21% of the more compact *Arabidopsis thaliana* genome (Ahmed et al. 2011) to 85% of maize genome or 70% of Norway spruce genome (Nystedt et al. 2013). The comparison of the TE content of genomes is not straightforward as the use of different methods to annotate TEs and the different quality of the reference genome, which for instance may include a variable fraction of unassembled reads corresponding to repetitive sequences, may introduce some biases.

Although not completely comparable, all the data obtained so far reveal the diversity and prevalence of TEs in plant genome. In order to know which regions of the genomic sequence correspond to TEs, the annotation and classification of these elements is required. Wicker et al. (2007) suggest a system to classify transposons based on their transposition mechanism (Class I via an RNA intermediate and Class II through a DNA intermediate), then into superfamilies according to coding region or TSD length, and finally into families according to sequence similarity. The most used criteria to place two sequences in the same family is sharing the 80% identity along 80% of sequence length (Wicker et al. 2007).

Nowadays several programs are widely distributed to analyze TEs in genome sequences (Bergman and Quesneville 2007). The most accurate way to annotate TEs is based on genome-specific consensuses which represents the most conserved sequence in a family which is then used to search for copies in the genome. The computational methods to discover these consensus elements, or representatives, can be classified in three types: *de novo*, homology-based and structure-based methods. *De novo* methods are based on clustering repetitive sequences without using prior information about TE structure or similarity to other known TEs. These methods allow to obtain consensus from new TE large families, but not for those with low copy number.

The methods based on similarity are designed to discover new TEs taking advantage of well-characterized TEs from databases, such as Repbase (Bao et al. 2015). This approach is possible because coding sequences tend to be well conserved in certain TEs, like RT of retrotransposons or TPase of DNA transposons (Wicker et al. 2007). So, more degenerated elements without coding capacity, such as MITEs, will not be identified by this approach.

The structure-based methods identify structural characteristics common in different TEs, such as LTRs or TIRs. Although low copy number families are detected, these approaches are limited to elements that share conserved structural characteristics. Specific bioinformatic tools must be designed to detect each type of TEs. Several tools have been developed to look for LTR retrotransposons based on identifying long direct repeats within a certain window in the genome, for example LTRharvest (Ellinghaus et al. 2008) or LTR_FINDER (Xu and Wang 2007). Also, MITEs can be identified by their structural features, such as length, TIRs and copy number, for instance using TRANSPO (Santiago et al. 2002) or MITE-Hunter (Han and Susan 2010).

Once the representatives are identified, the next step is to identify copies in genome sequence. This step is required because a TE family is also composed of fragments and degenerated elements which wouldn’t have been identified in the previous step.

There are two main purposes for annotating TEs in genomes. The TE annotation can be used only to mask the genome in order to more easily annotate the gene content, but it can also be used with the aim to study the biology and evolution of TEs (Bergman and Quesneville 2007).

There are different strategies and pipelines to perform global annotations useful for masking and to obtain a global genome TE annotation. One of the most comprehensive and used is the REPET package (Flutre et al. 2011). It has been used to annotate TEs in plant genomes, such as coffee (Denoeud et al. 2014), oak (Plomion et al. 2015) and black raspberry (VanBuren et al. 2016). REPET is composed by two pipelines, *TEdenovo* and *TEannot* (see Annex). Both use several programs in parallel and then combine their results to improve the exactitude and comprehensiveness (Permal et al. 2012). The *TEdenovo* pipeline generates consensus sequences using similarity, *de novo* and structural approaches based on repetitiveness, similarity and on structural data. Then those consensuses are used to find the copies in the genome using the *TEannot* pipeline.

Our group has collaborated with other research groups from our institute to sequence the melon genome (groups of Dr. Pere Puigdomènech and Dr. Jordi Garcia-Mas), and was in charge of the annotation and classification of transposable elements.

As a first approach, the strategy used to annotate TEs in melon genome was based on homology searches with known TEs complemented with an analysis of structural characteristics of particular TE families. In that way, a restricted but high quality annotation of young TE families was obtained, corresponding to a 19.7% of the melon genome space (Garcia-Mas et al. 2012). This TE annotation underestimates the real genome fraction that TEs occupy in melon due to its high stringency.

This project started with the idea to gain insight into melon TEs evolution. As cucumber is a close relative species, we wanted to compare the melon transposon landscape with that of cucumber to study the evolution of TEs since the divergence of these two *Cucumis* species.

Cucumber (*Cucumis sativus* L.) (2n=2x=14) and melon (*Cucumis melo* L.) (2n = 2x = 24) are two economically important vegetable crops and belong to the genus *Cucumis* (family Cucurbitaceae) (Schaefer 2007). Both species are of Asian origin and diverged from a common ancestor 10 Mya (Sebastian et al. 2010). The genome size of melon is approximately 450 Mb and that of cucumber is 367 Mb (Arumuganathan and Earle 1991), which is smaller than other species in Cucurbitaceae family. Cucumber is the *Cucumis* species with only seven chromosomes (Kirkbride 1993). Recent evidences suggest that cucumber evolved from a species with 12 chromosomes that shared a common ancestor with melon which were fused into its seven chromosomes (Huang et al. 2009; Li et al. 2011a).

Although melon and cucumber lineages split 10 Mya (Sebastian et al. 2010), the genome sequences of these two species are highly conserved. 62% of the cucumber genome aligns with melon (Sanseverino et al. 2015). Comparative mapping and sequence alignment studies showed a high syntenic relationship between melon and cucumber chromosomes (Li et al. 2011a; Yang et al. 2012; Yang et al. 2014). Despite the difference in size, both genomes have similar number of protein-coding genes (Li et al. 2011b; Garcia-Mas et al. 2012) suggesting that TEs have transposed and amplified to a greater extent in melon compared to cucumber (Garcia- Mas et al. 2012).

Indeed, a preliminary analysis based on the comparison of the melon stringent TE annotation and a rough TE annotation of cucumber TE families suggested that TEs have amplified to a greater extent in melon as compared to cucumber (Garcia-Mas et al. 2012). In order to perform a more reliable comparison, we decided to use the same tools and parameters to annotate TEs in melon and cucumber genomes and to study their TE landscape.

1.2.- OBJECTIVES

The main goal of the work described in this chapter is to study the TE landscape of the melon and cucumber genomes to understand better how TEs have shaped both genomes since the divergence of the two species. The specific objectives are:

- Annotate transposons in melon and in two cucumber genomes
- Study the distribution of TEs and genes across the genomes of these two *Cucumis* species
- Investigate the influence of TEs in chromosome structure and the evolution of genes

1.3.- MATERIAL AND METHODS

Genome sequences and gene annotation

The melon reference genome and gene annotation used was v3.5, available at http://www.melonomics.net (Garcia-Mas et al. 2012). This reference genome corresponds to a doubled-haploid line, named DHL92, obtained from a cross between Spanish variety ‘Piel de sapo’ (PS) and Korean accession ‘Songwhan Charmi’ (accession number: PI 161375).

Two different cucumber reference genomes and their gene annotations were downloaded online. The American Gy14 gynoecious inbred line data was obtained from Phytozome JGI database (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus), and the cucumber ‘Chinese Long’ variety from International Cucurbit Genomics Initiative (ICuGI) database (Huang et al. 2009; Li et al. 2011b).

Identification of transposable elements in genome

The REPET package v2.2 (Flutre et al. 2011; Quesneville et al. 2005) was used to search and annotated transposable elements within melon and cucumber genomes. The *TEdenovo* pipeline was run using default parameters, and step 8, corresponding to the clustering of consensus, was excluded. The *TEannot* pipeline was run using WU-BLAST 2.0 (Washington University – BLAST, http://blast.wustl.edu/), sensitivity for BLASTER of 2 (BLR_sensitivity: 2) and steps four and five, corresponding to the SSR detection, were excluded.

1.4.- RESULTS

Annotation of transposable elements

Transposable elements were annotated in melon and two cucumber reference genomes using the REPET package (Flutre et al. 2011), in order to have comparable TE annotations. In case of melon, TEs were annotated across the entire genome, including chromosome zero (chr00) which corresponds to all non-anchored scaffolds. A total of 168,008 transposable elements were identified, representing a 48% of genome space. TEs were classified when possible into the two major TE classes and we filtered out overlapping TEs from different classes. A total of 95% of the annotated transposon-related sequences were classified. The retrotransposon elements (Class I) account for 35.2% of the genome whereas DNA transposons (Class II) represent 8.35%, and few sequences left as unclassified TEs (0.06%) (Table 1.1).

Table 1.1. Comparison of the number of copies annotated as TE-related sequences and their fraction in the genome. TE confused correspond to those TEs couldn't be categorized in one of the two classes.

	Melon		Cucumber CL		Cucumber Gy14	
	# of copies	% of genome	# of copies	% of genome	# of copies	% of genome
class I	133232	35.20	79592	24.34	76090	25.59
class II	26009	8.35	7041	2.23	3075	1.21
TE confused	290	0.06	5	0.004	27	0.01
Total	159531	43.60	86638	26.58	79192	26.81

We compared the REPET annotation to the published one (Garcia-Mas et al. 2012). This new melon TE annotation is more comprehensive than the one published, and TEs account for almost the double of the genome space (from 19.4% up to 43.6%). Using this new TE annotation, the 70% of the genome is annotated either as gene or as TE (Figure 1.1a), considering 450 MB as the estimated genome size (Arumuganathan and Earle 1991). Each chromosome of melon genome has roughly the same percentage of annotated fraction. This wider-range TE annotation approach can be useful data for masking in order to annotate genes or studying the TE landscape of the genome.

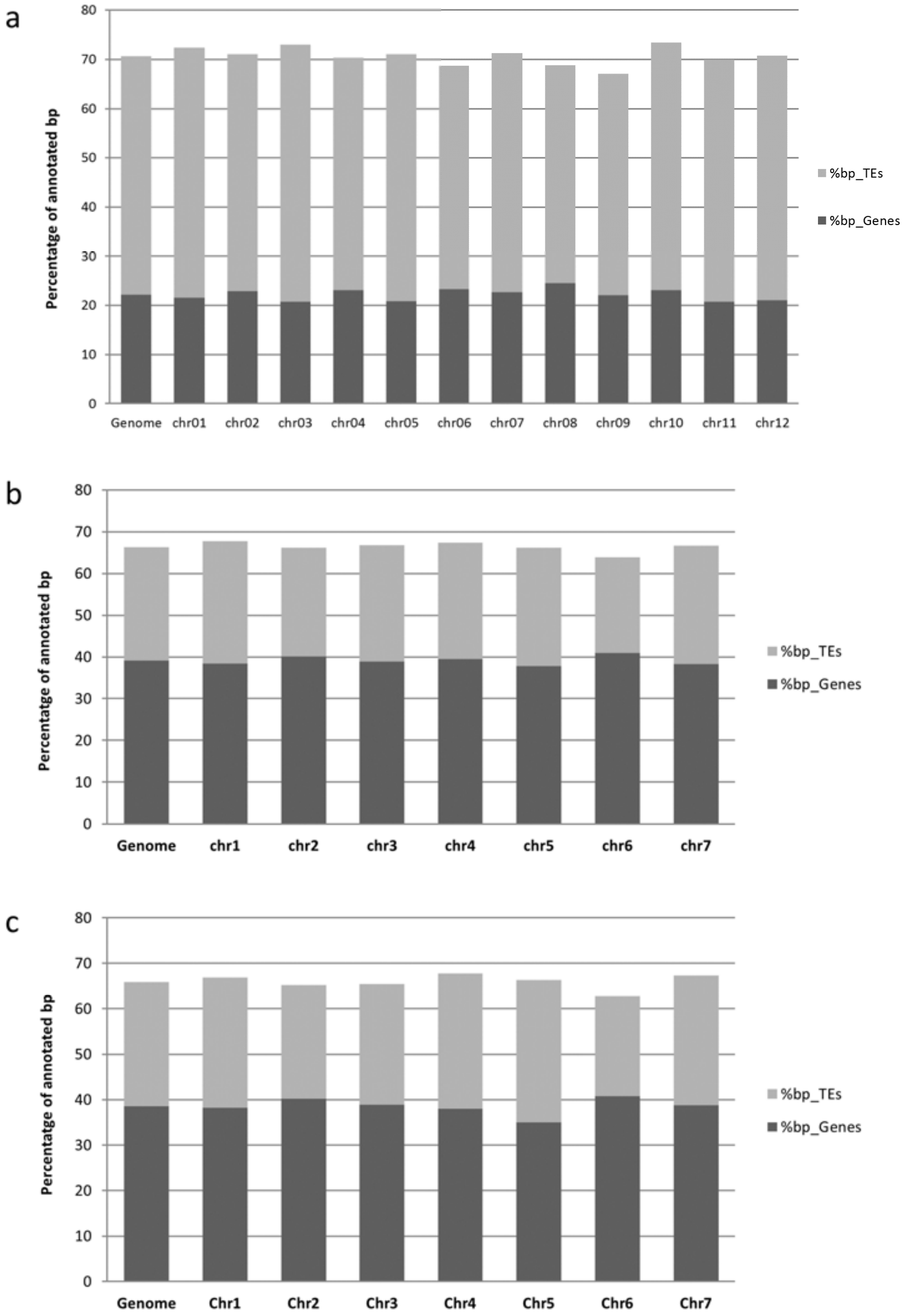


Figure 1.1. Percentage of annotated genes (dark grey) and transposable elements (light grey) per each chromosomes of the melon genome (a), the Gy14 cucumber genome (b) and the Chinese Long cucumber genome (c). The first column represents the percentage of annotated nucleotides in genome.

Five cucumber genomes have been sequenced and assembled by different consortiums or institutions, which are available on internet. The first one is the 'Chinese long' inbred line 9930 (Huang et al. 2009), which is commonly used in modern cucumber breeding. And the second one is the wild accession PI183967 (CG0002), which corresponds to *C. sativus* var. *hardwickii* accession (Qi et al. 2013). Both of them can be found in the Cucurbit Genomics Database (ICuGi, http://www.icugi.org/cgi-bin/ICuGI/index.cgi). The American Gy14 gynoecious inbred line was sequenced and assembled by Phytozome Join Genome Institute (JGI, http://jgi.doe.gov/). A polish Consortium of Cucumber Genome Sequencing was in charge of the north-european Borszczagowski variety (line B10) sequencing (Wóycicki et al. 2011). Finally, the Institute of Vegetable and Flowers from Chinese Academy of Agricultural Sciences (China, http://cucumber.genomics.org.cn/) has sequenced and assembled the genome of the domestic cucumber, *C. sativus* var *sativus* L.

We decided to work with the cucumber Gy14 and 'Chinese Long' genomes, because these genomes have a better assembly. We used the cucumber Gy14 genome from Phytozome, which is a commercial gynoecious inbred line. Dr. Jordi Morata, a member of the group, anchored 235 scaffolds using a high-density genetic map (Yang et al. 2012) in order to obtain the seven cucumber pseudomolecules. REPET was run using all scaffolds greater than 10 kb to annotate TEs. Then, Dr. Morata transferred the gene and TE annotations from scaffolds to chromosomes.

The other cucumber genome analyzed in this study corresponds to variety commonly used in modern cucumber breeding, named 'Chinese Long' (CL). The chromosomes can be downloaded from ICuGI website, as well as its gene annotation (http://www.icugi.org/cgi-bin/ICuGI/index.cgi). The TE annotation was obtained using the REPET package on the seven chromosome sequences.

The estimated genome size of cucumber is 367 Mb size (Arumuganathan and Earle 1991), and the genome assembly size is about 191 Mb and 171 Mb in CL and Gy14 cucumber genomes, respectively. So, around 50 to 60% of estimated genome size has been assembled, whereas melon corresponds up to 83% (450 Mb estimated genome size and 375 Mb assembled).

This fact could generate a bias, because we could be losing an important TE fraction from cucumber genomes. We checked the percentage of annotated TEs in the unanchored genome fraction which is up to 43.4% in melon and 32.74% in cucumber Gy14. Both of them are in line with the proportion of annotated TEs in the whole genome (32.74% in melon and 27.12% in Gy14), meaning that similar proportion would not be taken into account in both genomes.

In both cucumber genomes, TEs were annotated using the same strategy as in melon. And also, overlapping TEs from different classes were filtered out, in this case 0.5% and 1.5% for Gy14 and CL, respectively (Table 1.1). Around a quarter of the genome space corresponds to retrotransposon elements (Class I) in both cucumber genomes whereas DNA transposons (Class II) account for 1.21% in Gy14 line and 2.23% in CL line (Table 1.1). The 65-66% of the cucumber genomes is annotated either as gene or as TE (Figure 1.1b and 1.1c), and those levels are roughly maintained in all chromosomes for both cucumber genomes.

Distribution of genes and TEs across genome

In order to investigate the distribution of genes and TEs in melon and cucumber genomes, the densities of TEs and genes in each chromosome were plotted. Gene and TE distributions show an anti-correlation, where TEs are not homogeneously distributed along chromosomes. In the case of melon, we can define two types of regions: TE-rich and gene-rich regions, based on the high or low density of TEs and genes per 1Mb bin (Figure 1.2). One can observe that TE-rich regions account for the majority of the chromosomes, including pericentromeric regions, while gene-rich regions are small and, most of the cases, restricted to the chromosome arms close to the telomeres. This was observed in a preliminary analysis of melon TEs (Garcia-Mas et al. 2012) but has been confirmed with the new TE annotation.

The chromosomal distribution of these two rich regions matches perfectly to the distribution of recombination rate (Garcia-Mas et al. 2012), indicating that recombining regions coincide with gene-rich regions and TE-rich regions correspond low recombining regions around the centromeric regions.

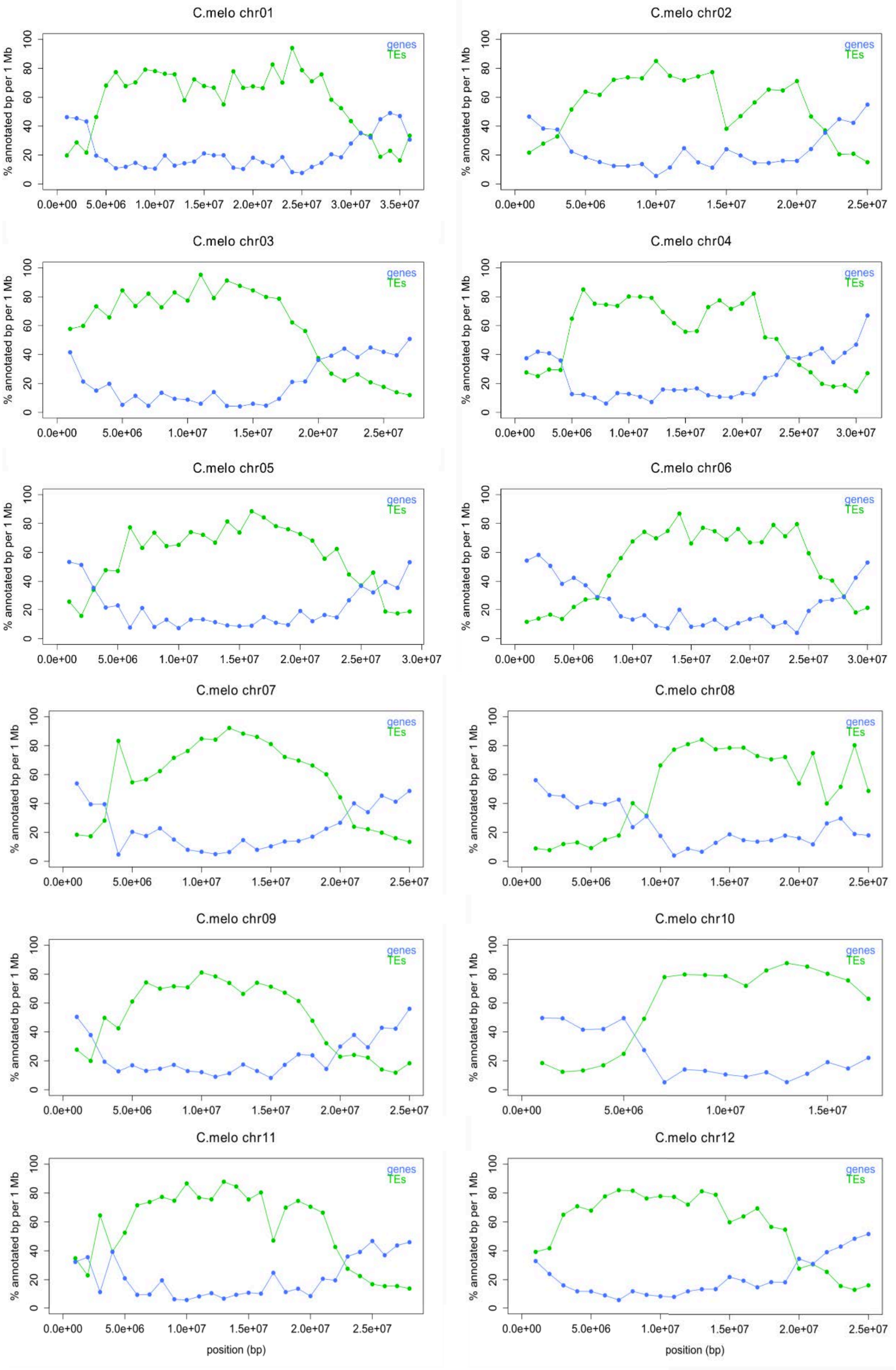
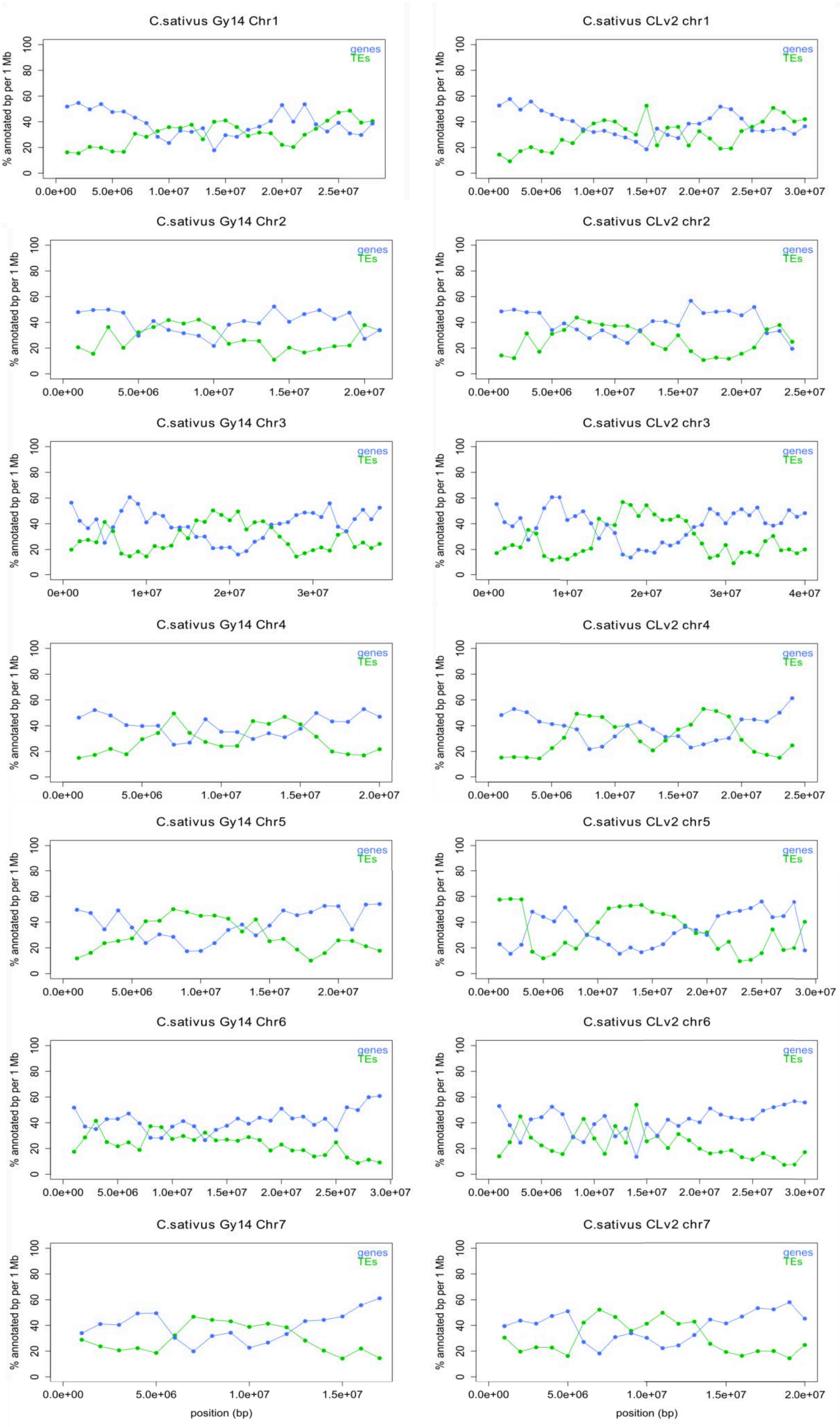


Figure 1.2. Distribution of TEs (green) and genes (blue) for each of the 12 chromosomes in the melon genome. Each point corresponds to the percentage of nucleotides annotated as a gene or TE in a 1Mb window.


C.sativus Gy14 Chr1
C.sativus CLv2 chr1
C.sativus Gy14 Chr2
C.sativus CLv2 chr2
C.sativus Gy14 Chr3
C.sativus CLv2 chr3
C.sativus Gy14 Chr4
C.sativus CLv2 chr4
C.sativus Gy14 Chr5
C.sativus CLv2 chr5
C.sativus Gy14 Chr6
C.sativus CLv2 chr6
C.sativus Gy14 Chr7
C.sativus CLv2 chr7
genes
TEs
% annotated bp per 1 Mb
0
20
40
60
80
100
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
3.0e+07
0e+00
1e+07
2e+07
3e+07
4e+07
position (bp)

Previous page:

Figure 1.3. Distribution of TEs (green) and genes (blue) for each of the 7 cucumber chromosomes, Gy14 in left column and 'Chinese Long' in right side. Each point corresponds to the percentage of nucleotides annotated as a gene or TE in a 1Mb window.

The analysis of the distribution of TEs and genes along cucumber chromosomes, both in Gy14 and 'Chinese Long', shows that there is no single long TE-rich region in most chromosomes like in melon. These TE-rich regions are quite small and are interrupted in some cases by small gene-rich regions (Figure 1.3). For example, chromosome 4 has two clear TE-rich regions flanked by two gene-rich regions, but this TE-rich region is interrupted by a gene-rich region in the middle. Comparing the distribution of these two cucumber genomes, most of chromosomes present the same profile, and differences may be due to quality of genome and the proper orientation of anchored scaffolds.

When we compare the TE content in melon and cucumber genomes, TEs have amplified to a greater extent in melon accounting for 43% in this species whereas they account for only 26% in cucumber.

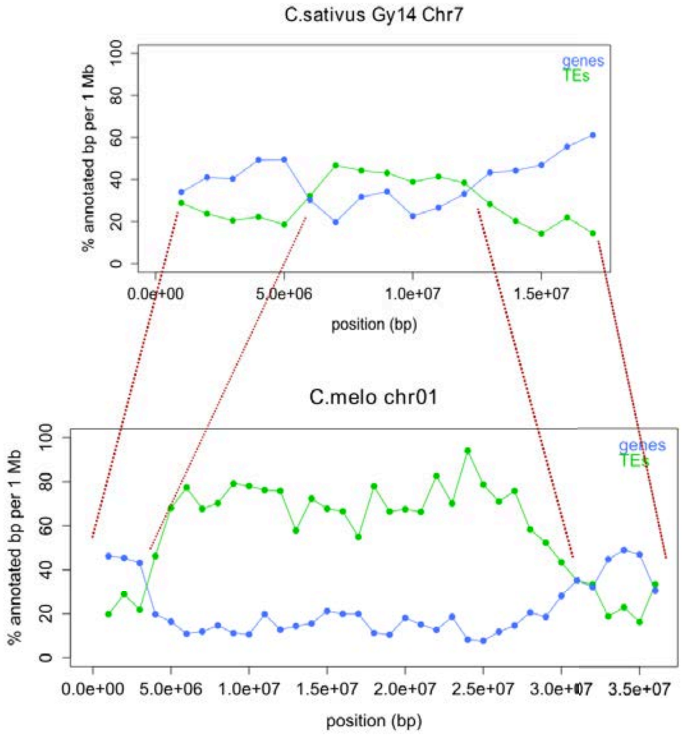


Figure 1.4. Distribution of TEs (green) and genes (blue) in melon chromosome 1 and Gy14 cucumber chromosome 7. Each point corresponds to the percentage of nucleotides annotated as a gene or TE in a 1Mb window. Syntenic regions of a same size are indicated by red lines.

Due to high collinearity across the entire chromosomes (Li et al. 2011a; Yang et al. 2012; Yang et al. 2014), we compare melon chr1 versus cucumber Gy14 chr7, noticing that pericentromeric TE-rich region has expanded in melon (Figure 1.4). This is linked to the greater TE activity and the absence of recombination in the pericentromeric TE-rich region in melon (Garcia-Mas et al. 2012). Whereas this region is much bigger in melon compared to cucumber, the gene-rich regions are the same size, approximately.

Interestingly, whereas in melon recombination is high in gene-rich regions and is almost completely suppressed in the long TE-rich pericentromeric regions, in cucumber the recombination rate is essentially constant along chromosomes (Figure 1.4).

Impact of transposable elements in gene evolution in melon and cucumber

The results presented above show that TEs have expanded pericentromeric TE-rich regions in melon and that recombination rate is suppressed in melon pericentromeric regions. On the contrary in cucumber, where TEs have not expanded the pericentromeric regions, recombination is constant along chromosomes.

Recombination allows for the independent evolution of loci, and therefore differences in recombination rates may affect profoundly gene evolution. For this reason, we have decided to analyze the possible differences in the evolution of melon and cucumber genes, and in particular melon genes sitting in pericentromeric regions.

Due to the collinearity between melon chr1 and cucumber chr7, orthologous genes can be identified to perform a preliminary study on whether genes melon/cucumber orthologous genes located in pericentromeric TE-rich regions (low recombination and expanded regions in melon) have evolved differently compared to those located in gene-rich regions (high recombination rate). Regions were defined as TE-rich when the percentage of annotated TEs was greater than that of genes, whereas the rest of the chromosome was considered as gene-rich.

As already explained, TEs have not accumulated to a high extend in cucumber, and most of genes in chr7 are located in gene-rich regions. On the contrary in melon, genes are equally distributed among gene-rich and TE-rich regions (Table 1.2).

Table 1.2. Number of genes located in the two different rich regions, when comparing entire collinear chromosomes.

	Number of genes		
	Total	**Gene-rich**	**TE-rich**
C. sativus chr7	2107	1341 (63%)	766 (37%)
C. melo chr1	2515	1191 (47%)	1324 (53%)

Having as an example two collinear chromosomes, we focused on analyzing orthologous genes from melon and cucumber and identifying when they were located in equivalent regions with respect to the TE richness regions or not. Dr. Morata identified orthologous 1-to-1 gene pairs of cucumber and melon, meaning that one melon gene has only one orthologous gene in cucumber and vice versa.

Table 3 shows that while most of orthologous genes are located in equivalent regions, around 180 orthologous genes are located in gene-rich regions in cucumber and TE-rich regions in melon, and no genes present the reverse distribution.

Table 1.3. Number of orthologous genes found in each rich region. The rate corresponds to the number of melon genes versus cucumber genes.

		C. sativus	
		Gene-rich	**TE-rich**
C. melo	**Gene-rich**	711 / 684	0 / 0
	TE-rich	188 / 180	303 / 282

This is in line with the fact that TEs have expanded pericentromeric TE-rich regions in melon and the number of genes in TE-rich regions is higher in melon compare to cucumber. These preliminary results have prompted us to analyze the evolution of melon genes orthologous or non-orthologous to cucumber genes at a global scale.

This ongoing project suggests that TEs have shaped the evolution of melon genes, as they seem to evolve differently in the two chromosomal compartments which have been defined depending on TE-richness, and may have consequences in the evolution of the two species.

1.5.- DISCUSSION

Transposable elements are mobile genetic elements that accumulate as repetitive sequences of different length and structure in genomes. There are several computational tools which have been developed to detect and annotate TEs in assembled genomes. These programs follow different approaches: detection by similarity to other TEs, by structure of TEs or by their repetitiveness.

The difficulties of TE annotation may vary among genomes for different reasons. On the one hand, genomes may have different TE activity patterns (Hoen et al. 2015). If a genome has low TE activity, most TE copies may come from ancient bursts of transposition from just a few TE families (Kazazian 2004). Over time, those copies accumulate mutations differing from the original sequence and making difficult their detection. But also in cases of recent TE activity in genomes their annotation is still challenging, for example when complex structures carried by internal deletions or nested insertions are present (Quesneville et al. 2005). In addition, some TEs may be present at low copy numbers, and others, in particular non-autonomous short elements such as MITEs, may contain very few structural characteristics easy to recognize, making their annotation not straightforward.

As the specificity and sensitivity of different methods used for TE annotation is different, when comparing the TE content of different genomes the use of the same method and parameters is crucial (Hoen et al. 2015). Depending on the objective, more or less stringent TE annotation can be used. For example, for evolutionary or phylogenetic studies a stringent annotation is more useful, while a less-conservative annotation is able to detect more degenerate elements and this approach can be useful for masking genomes or study TE landscape.

For instance, Maumus et al. 2014 combined several repeat annotation programs and increased at least 20% the genome covered by TE annotations in *A. thaliana* genome compared to annotations obtained from a single program. The challenge is where to put the cutoffs and thresholds and which program to use in order to obtain a TE annotation suitable for each purpose. Among all pipelines for TE annotation available, the REPET package (Flutre et al. 2011) is the most common used one to detect, annotate and analyze repeats in genomic sequences.

In this work, we wanted to compare the TE content and characteristics of melon and cucumber using REPET with the same parameters in order to allow a consistent comparison. However, it has to be noted that the available genome sequence of melon and cucumber differs in their assembly quality, which may also affect TE annotation. The first TE melon annotation (Garcia-Mas et al. 2012) was a conservative annotation, where the most abundant and recent TEs were annotated. We now used the REPET package to generate a more in-depth genome-wide annotation of TEs. The similarity and structural approaches that uses REPET allowed us to annotate as TE-related sequences up to 43% of melon genome, whereas we detected TEs covering only the 26% of the cucumber genomes (Table 1.1). The higher TE content of the melon genome is in line with its bigger size as compared with cucumber. Indeed, although not only directly comparable for the reasons already explained, there is a relationship between TE content and genome size. Whereas the 21% of *A. thaliana* genome (120 Mb) corresponds to TEs (Ahmed et al. 2011), tomato genome (900 Mb) is around 63% (Tomato Genome Consortium 2012). The same tendency occurs in the two analyzed *Cucumis* species: 26% of cucumber genome (367 Mb) and 43% of melon genome (450 Mb).

The comparison of TEs in the melon and cucumber genomes has provided new insight of their evolution. Whereas melon and cucumber are closely related species, with a divergence time of some 10 Mya (Sebastian et al. 2010), no common transposons have been found (data not shown). This is not surprising as noncoding sequences evolve faster than coding sequences (Freeling et al. 2012), and melon and cucumber genomes can only be aligned over their coding sequences.

These two closely related genomes have large syntenic chromosomal regions (Li et al. 2011a; Yang et al. 2012; Yang et al. 2014). For example, most melon chromosome 1 is syntenic to cucumber chromosome 7. When comparing these two largely syntenic chromosomes it can be clearly seen that whereas the euchromatic chromosomal regions of both chromosomes span a region of similar length, and seem to have accumulated a similar amount of TEs, the pericentromeric regions in the melon chromosome are much larger due to the accumulation of more TEs. This suggests that the higher TE activity in melon has resulted in an increase of pericentromeric regions in this species and shows that TEs can have an important impact in shaping the structure of chromosomes.

The chromosomal distribution shows anti-correlation between TEs and genes in both genomes. TEs tend to concentrate in pericentromeric zone, although this region is much bigger in melon. Interestingly, the recombination rate is uniform across cucumber chromosomes (Lou et al. 2013), while in melon the recombination is negatively correlated with TE density and is greatly reduced in the pericentromeric regions. This suppression of recombination in the TE-rich pericentromeric regions has been observed in many other species where TEs are concentrated in pericentromeric regions, such as tomato (Tomato Genome Consortium 2012) and wheat (Choulet el al. 2014).

Recombination is in general repressed in repetitive regions, which show heterochromatic epigenetic marks (Zamudio et al. 2015), to avoid pairing of non-allelic positions that may lead to meiotic problems. But at the same time the regions that present a low recombination rate may accumulate TEs as they are more difficult to be eliminated by selection (Gaut et al. 2007). In any case, the increase in size of the TE-rich pericentromeric region in melon may have changed the recombination frequency distribution along the chromosomes, as it has been recently proposed for *Arabis alpina* with respect to *A. thaliana* and *A. lyrata* (Willing et al. 2016).

Although these regions are highly TE-rich, they also contain some genes. The fact that recombination rate is lower in these regions may affect how genes evolve. The syntenic chromosomes 1 of melon and 7 of cucumber allowed us to obtain a list of orthologous genes which are located in similar or different compartments with respect to the TE-richness and analyze their evolution. No functional enrichment has found from genes located in the two different regions, and other approaches should be undertaken. But for instance, the study of TE distribution in tomato genome revealed that DNA TEs are associated with genes differentially expressed during fruit ripening (Jouffroy et al. 2016).

The comparison of melon and cucumber TE landscapes is an undergoing project. A global analysis of these two genomes, carried by other lab members, suggest that the pericentromeric TE-rich regions of melon concentrate more melon specific genes.

Finally, the chromosomal distribution of TEs in melon makes it an interesting species to study how genome maintains TEs under control. TE insertions are generally deleterious and the ones inserted close to genes are selected against (Hollister and Gaut 2009). To get an idea of the impact of TEs, it could be interesting to investigate polymorphisms due to TEs in a wide range of varieties (see Chapter 2).

CHAPTER 2

Impact of transposon insertion variability during breeding and domestication processes

CHAPTER 2: Impact of transposon insertion variability during breeding and domestication processes

Analysis of transposon insertion variability in different varieties of melon, date palm and *Physcomitrella patens*

2.1.- INTRODUCTION

Transposable elements are an important source of genetic variability useful for evolution both for wild and domesticated plants (Lisch 2013; Olsen and Wendel 2013). Examples of TE insertions which cause phenotypic changes to important agronomic traits include the different skin colors in grapevine (This et al. 2007), the nectarine phenotype in peaches (Vendramin et al. 2014), the different flesh fruit color in blood orange (Butelli et al. 2012), the seedless phenotype in apples (Yao et al. 2001) and the sex determination in melon flowers (Martin et al. 2009).

Besides the examples listed above, there is a lack of understanding how TEs have impact in the evolution of eukaryote genomes. The availability of sequenced genomes provides information about the TE content and its distribution across the genome. But, in order to understand the role of TEs during evolution of a certain species, resequencing different varieties from the same species or close relatives is required.

The new era of genomics allows generating a huge amount of resequencing data of many different crop varieties and landraces. For instance, 3000 rice varieties have been sequenced which allow to study the genetic bases of many different important traits (Li et al. 2014). However, the resequencing data of varieties have been analyzed to investigate genetic variation, such as SNPs, small InDels, PAVs, but have not been widely used to analyze TE insertion polymorphism. The reason for that is that the analysis of TE polymorphisms at genome-wide scale is not straightforward. To date, there are several tools for detecting TE insertion polymorphism using paired-end resequencing data, such as VariationHunter (Hormozdiari et al. 2010), Retroseq (Keane et al. 2013), ITIS (Jiang et al. 2015) and T-lex2 (Fiston-Lavier et al. 2015).

In our research group of *Structure and evolution of plant genomes* (CRAG, Barcelona), we have developed a new bioinformatic tool to study TE insertion polymorphism. Indeed, in collaboration with the group of Dr. Stephan Ossowski (CRG, Barcelona), we developed the program Jitterbug (Hénaff et al. 2015) that allows detecting TE insertions comparing the paired-end sequencing of varieties to a reference genome.

The purpose of this chapter is to analyze the impact of transposition on gene and genome evolution in melon (*Cucumis melo*), date palm (*Phoenix dactylifera*) and the moss *Physcomitrella patens*. To this end, we used the Jitterbug program, as well as other available programs, to identify polymorphisms among plant varieties due to the presence or absence of transposable element at a given loci using paired-end sequencing data. This analysis can provide information about the TE activity in the evolution of these plant genomes and to what extent TEs have impacted during their evolution.

2.2.- OBJECTIVES

This chapter consists in studying the contribution of transposons to recent evolution in three species: melon, date palm and *Physcomitrella patens*.

- Identify polymorphisms in different varieties due to the presence or absence of transposable element at a given loci
- Analyze the impact of PM-TEs on gene and genome evolution

2.3.- MATERIAL AND METHODS

Resequencing data

a) Melon

Seven melon accessions were analyzed in this study (Table 2.1). Paired-end libraries, having 500bp fragment length and 150 read length, were produced and sequence using Illumina Genome Analyzer IIx technology at the Centre Nacional d'Anàlisi Genòmica (CNAG, Barcelona). The melon reference genome corresponds to a doubled-haploid line, named DHL92, obtained from a cross between PS and SC (Garcia-Mas et al. 2012).

Table 2.1. Melon accessions used in this study

Plant designation	Accesion number	Code	Cultivar group	Subspecie	Origin	Reference
DHL92[1]	DHL92	DHL92				Garcia-Mas et al. 2012
Piel de sapo	T111	PS	*Inodorus*	*melo*	Spain	Garcia-Mas et al. 2012
Songwhan charmi	PI 161375	SC	*Conomon*	*agrestis*	Korea	Garcia-Mas et al. 2012
Cabo Verde[2]	C-836	CV		*agrestis*	Cabo Verde	Gonzalez et al. 2013
Calcuta	PI 124112	CAL	*Momordica*	*agrestis*	India	Sanseverino et al. 2015
Irak	C-1012	IRK	*Dudaim*	*melo*	Irak	Gonzalez et al. 2013
Trigonus[2]	Ames 24297	TRI		*agrestis*	India	Sanseverino et al. 2015
Vedrantais		VED	*Cantaloupensis*	*melo*	France	Sanseverino et al. 2015

[1]DHL92 is a doubled haploid line derived from PI 161375 x T111 and represents the melon reference genome
[2]Unknown cultivar group

b) Date palm

The BAM files of 69 date palm varieties were provided by Dr. Khaled Michel Hazzouri (NYU, Abu Dhabi, United Arab Emirates), as well as the reference genome (Al-Mssallem et al. 2013) and annotations of both genes and transposable elements.

c) *Physcomitrella patens*

The paired-end reads were downloaded from SRA database of NCBI with SRP004339 as project code (Experiment Numbers: SRX030894 and SRX037761). Both experiments correspond to *P.patens* var. Villersexel libraries, having 3 runs for SRX030894 with a total of 277.9 M reads (42.2 Gbases) and 2 runs for SRX037761 with a total of 176.1 M reads (26.8 Gbases), respectively (Table 2.2).

Table 2.2. Characteristics of SRA runs from the project number SRP004339 with coverage and standard deviation of each bam file

Experiment Code	Run Code	# of Spots	# of Bases (G)	file size (Gb)	Coverage bam file (x)	StDev bam file
SRX037761	SRR090654	29249996	4.4	76	4	7.9
	SRR512764	146861614	2.,3	17.4	16.08	31.31
SRX030894	SRR072296	41906798	6.4	40.4	8.62	16.72
	SRR191864	201288783	30.6	13.3	15.27	28.87
	SRR400524	34736515	5.3	2.3	8.35	15.92

Processing the reads

Reads were converted to FastQ format generating 2 files (one for forward reads and the other for reverse reads) using *fastq-dump* from SRA-Toolkit v2.4.4. Reads were trimmed and filtered using SGA (https://github.com/jts/sga) with *preprocess* (-q 10 –m 50 –permute-ambiguous –pre-mode=1) keeping read pairs untouched, also if a read should be discarded due to the read quality or reads with ambiguous base calls. Then, reads were indexed and corrected using SGA *index* (-d 2000000) and SGA *correct*, both with default parameters.

Corrected reads were mapped to the assembled reference genome using *align* and *sampe* from BWA program (Li and Durbin 2009). SAMtools v 0.1.18 was used (Li et al. 2009) to sort and index all *bam* files. Also, it was used to merge *bam* files.

Transposon polymorphism predictions

Jitterbug (Hénaff et al. 2015) (http://sourceforge.net/projects/jitterbug/?source=directory) was used to detect transposon insertions in the resequenced samples respect to the corresponding reference genome. Jitterbug was run with a value of 35 for the minimal mapping quality of the reads (-q 35). The predicted insertions were filtered using the calculated parameters specific for each library and, taking into account the consistency of the predicted inserted element comparing forward and reverse cluster reads.

Analyses to detect deletions were performed using Pindel software (Ye et al. 2009) in the resequenced samples respect to the corresponding reference genome. Pindel v2.4 was run with default settings, except for maximum range index (5) and anchor quality (35). In order to decrease computational time, report of inversions and duplications were disabled. The predicted deletions were filtered by size, selecting those greater than 200 bp and less than 25 kb. Then, additional filtering step was applied in order to select only these predicted deletions that overlap with annotated transposons. Intersections and manipulation of data were performed with Bedtools v2.17 (Quinlan and Hall 2010).

PCR validation of TE insertion polymorphism

A subset of the predicted TE insertion polymorphisms in seven melon varieties was analyzed by PCR. PCR products were obtained in a final volume of 20ul containing 40ng genomic DNA, 300mMdNTPs, 20mM for each primer, and 2 units/20ul of LongAmp Taq DNA Polymerase (New England BioLabs). Primer pairs were designed to be 20–26bp long for PCR amplification using Primer3 software (Untergasser et al. 2012). The oligonucleotides used are listed in Table 2.3. Half of the PCR products were separated on a 1% agarose gel and stained with ethidium bromide for checking the PCR amplification. Fragment sizes were estimated with the 1 kb DNA ladder (Biotools).

Melon candidate gene analysis

GO terms overrepresentation analysis was performed using the Cytoscape plugin BiNGO (Maere et al. 2005) to analyze the genes with polymorphic TE insertion between VED and PS.

To identify an orthologous of melon gene candidates in *Arabidopsis thaliana*, tBLASTx analysis was performed. The gene annotation of *Arabidopsis thaliana* was version TAIR 9 taken from www.arabidopsis.org.

Melon flowers material

Seeds from melon varieties ‘Piel de sapo’ (PS) and ‘Védrentais’ (VED) were kindly provided by the group of Dr. Jordi Garcia-Mas (CRAG-IRTA). Plants were grown in greenhouse (diurnal temperature 28ºC ± 2ºC, nocturnal temperature 22ºC ± 2ºC, relative humidity 60C ± 5ºC) and flowers at different stages of development were collected.

Table 2.3. List of oligos used in melon studies

Primer Name	Sequence (5' - 3' orientation)
CM_11055	GGAGCAAAGGAACTGAGAAAGA
CM_11055-R	ATACCTCATGCAGGAATTGGTAAT
CM_11174	GGTTTGATCTGAACCAATAAATCG
CM_11174-R	GCGTTGGAGGAAATAGAGAGATAA
CM_11998	CTCACCAATTCACTAAGCTCCA
CM_11998-R	CACCAAAGCCATGAGGAACTA
CM_1654	TTCATACCCAGCTCAAACCTCT
CM_1654-R	CACAATGTCACAACTCACATGC
CM_20546	AACTGTAAGAAGGAACGAAGAGGA
CM_20546-R	GATTCCTCACTCCAACAGTTGAC
CM_20740	TCTTCTAATTGCCTTCTCCACAG
CM_20740-R	GTTAAAGAATCGGAATCGTGTTG
CM_21258	AACTCCGCATGTTCTTGAGCT
CM_21258-R	TAGGTAGGTGACCATCATGGATT
CM_22554	GTAGCAGTACGCTGTTTCAACACT
CM_22554-R	TACACCCCTTGTGTCATTTATACG
CM_2374	CTTGGAGTCTATGAATGGAGTGG
CM_2374-R	GAGATAAAACTATGGGTGTGATTGG
CM_24088	CTACCAGCACAACCAACAACATA
CM_24088-R	GATCATCCGAAGTTTAAGAGAGGA
CM_2764	TCCCTTCCCTTACTCCAAATCTA
CM_2764-R	ACAATGTTGACAAGGAGATGACAC
CM_3698	TGTTCTACACCAACAGGGTCAC
CM_3698-R	TCTTTTCTAGGGATGTGACTAATCG
CM_4552	CGATGACTCCAATCTTATTCAGG
CM_4552-R	AAAGTTGTTCTTCACCAACAGGA
CM_4946	CAATGAGCAAAATGAAGGCATA
CM_4946-R	TACTCAAGAGTGTGTTCCTTTCCA
CM_6853	TGCAATTTCCGTAGTAACATTTG
CM_6853-R	GTAGGTTGGGGTTAGGAAGTCAC
CM_7125	ACTGATCCCAAGAACTCTGCTC
CM_7125-R	ACTAACCATACCCCGTTGATCTT
CM_7126	GCAAGTGACGAATGATGTCTGT
CM_7126-R	GGGACATACTTTGCGAGTAGATG
CM_8260	GATGAAACTGGAGGGATTAGAGG
CM_8260-R	TTCCAACTACATTGTTAGCGAGAG
CM_9115	CTCTTCCATCAAACCACCAGTAG
CM_9115-R	CCACAAGTGAGGAGGAGTGTTAG
CM_9716	TTTGATACTGCAACCTTGGTCGT
CM_9716-R	AACACTGCCAGTTGTCAAGTTAAG
CM_9728	CAACCCCATAGATGAGATGACA
CM_9728-R	GCAACTATCCACCCTTCAATACTT
CMins3_con	ACGTGACAAGGGACCGTAAA
CMins3_con-R	AACAAGAACCGCAAAACACC
CMins3_right_in_rc-R	TGAAGATCGAAGACCACGAA
CMins5_con	GCTTCCTCCACCTAGGCTCT
CMins5_con-R	CAAATTGGCACGCCTAGTAAG
CMins5_right_in_rc-R	CGACAAGGAATCTGCAACAG

Primers use	Primer Name	Sequence (5' - 3' orientation)
TE insertion validation of candidate genes	CM_cd1-fw	TCACCACATTGTGCTCCTTTC
	CM_cd1-rev	GACCTACATCGGCTTTCTTGTC
	CM_cd2-fw	AGGGTAATGGGCAGATAGCATA
	CM_cd2-rev	TGCAACACAACTCACCCATT
	CM_cd3-fw	ACTCCTTGTCAGACTTTTCATGTG
	CM_cd3-rev	TAGACGAAGCCATCCATTACCT
	CM_cd4-fw	GGCTCAAATGCCTTACAAGC
	CM_cd4-rev	CATGGAGAAATGGACTTGATGA
	CM_cd5-fw	TTCAATAAACGGCAGCCTCT
	CM_cd5-rev	ATGCCTGGTTCTTCGTACCTT
	CM_cd6-fw	AGCATGATTCCACTTTGTTGG
	CM_cd6-rev	CTGCTGCGTAAGCCATCTATC
	CM_cd7-fw	GCATTGACAGTGATGACATGG
	CM_cd7-rev	GGTATGGCTGCTGAATGTGTT
	CM_cd8-fw	GAGTTTCGCATCTGTTCTTATGG
	CM_cd8-rev	TGTCCAAATCGAAGATCAATAG
qRT-PCR analysis of candidate genes	CM_cd3_ex-fw	TTGGGAATTGAGGAAAGTGG
	CM_cd3_ex-rev	GTAGCGGCAAGGCATAGAAG
	CM_cd4_ex-fw	GCCAAGTTCAATCCAACGAT
	CM_cd4_ex-rev	CTCACTGGACTGGGGATAACA
	CM_cd7_ex-fw	GGTAGGGGACGGTGGATTAT
	CM_cd7_ex-rev	CAGGTCACCAGCAAGAACAA
	CM_cd8_ex-fw	TGGGTTTCTTATCTCCTTCCAA
	CM_cd8_ex-rev	AGATGACTGCTCTCCCCAGA

Expression analysis of melon candidate genes

RNA from melon flowers was isolated using PureLinkTM RNA Mini Kit (Applied Biosystems, Ambion). The isolated RNA was treated with DNA-freeTM kit (Applied Biosystems, Ambion), following the manufacturer's protocol. For assessing RNA quality and integrity, samples were visualized in agarose gel stained with ethidium bromide and quantified using a Nano-Drop ND-1000 (Thermo Scientific). The first strand of complementary DNA was synthesized from DNase-treated total RNA (1ug) with oligo(dT)18 and the SuperScriptTM III Reverse Transcriptase kit (Invitrogen), according to manufacturer's instructions.

The quantitative real-time PCR reactions (qPCR) were performed on optical 96-well plates in the Roche Light Cycler 480 instrument using SYBR Green I Master (Roche Applied Science), primers at 10uM and 20ng of total RNA, running each sample in triplicate. Cycling conditions were: 95ºC for 5 min (holding stage); then 95ºC for 10 seconds, 56ºC for 10 seconds and 72ºC for 10 seconds (amplification stage); and finally, the qPCR specificity was checked with the melting curve. Reverse transcriptase negative controls and non-template controls were systematically included. The housekeeping gene *CYCLOPHILIN7* (*CmCYP7*) (Saladié et al. 2015) was used to normalize the qPCR output, where Ct values of 40 or above were considered negative values or lack of amplification. Primers were designed using Primer3 software tool (Untergasser et al. 2012) and are listed in Table 2.3.

2.4.- RESULTS

In order to analyze the role of TE insertion polymorphisms, a combination of tools have been used to identify transposon insertions and deletions in the resequenced varieties with respect to the reference genome of the three species: melon, date palm and *P. patens*.

The "insertion" or "deletion" terms do not define the evolutionary process of transposon-related insertions and excisions, they only refer to the relative presence/absence of a TE in a particular resequenced variety with respect to the reference genome.

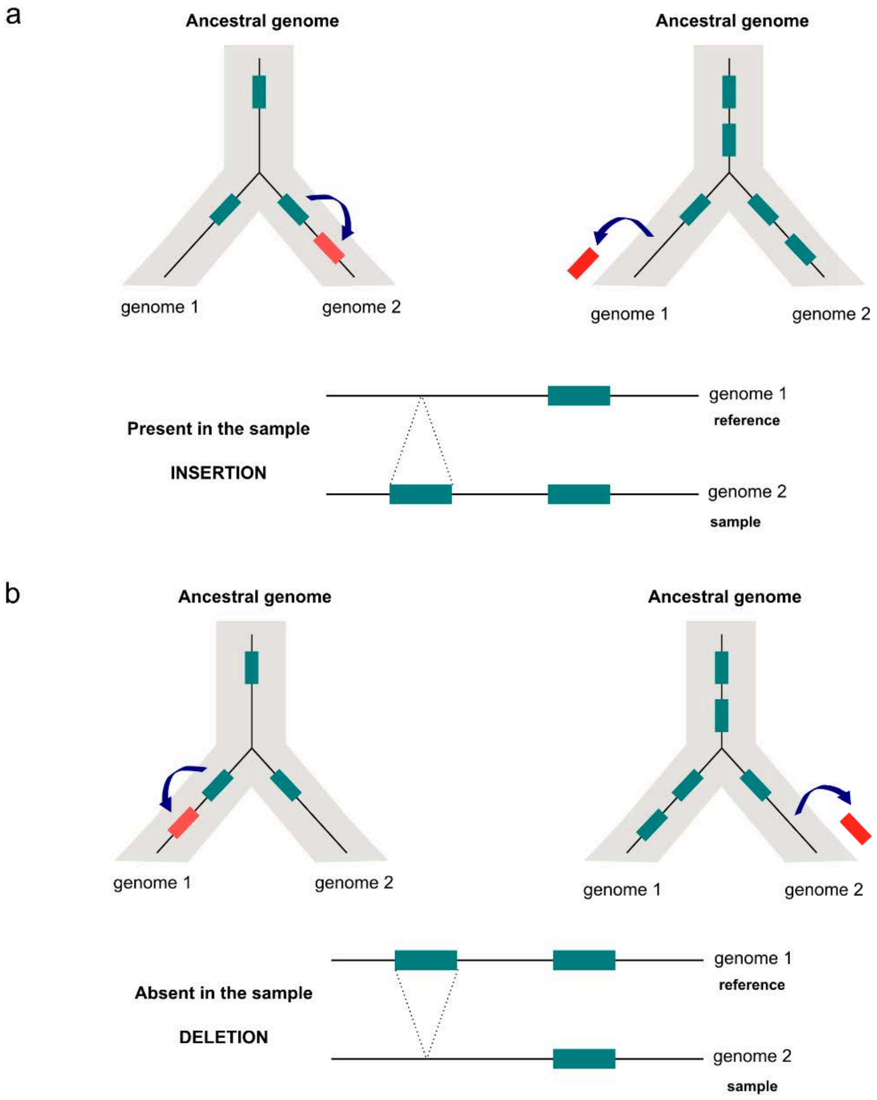


Figure 2.1. The presence or absence of TE at given locus is not linked to transposition events. When comparing a re-sequenced sample to the reference genome, insertion corresponds to the presence of a TE in the sample (a) and deletion absence in the sample (b).

For instance, a detected insertion can be either a TE deleted in the reference or a TE inserted in the sample (Figure 2.1a), while a deletion can be either a TE inserted in the reference or a TE deleted in the sample (Figure 2.1b). Detecting "absences" in the sample is quite straightforward and there already exist several programs which detect the absence of any type of sequence. We evaluate Pindel (Ye et al. 2009) and Breakdancer (Chen et al. 2009) with our data, and we chose Pindel because of better sensitivity (Sanseverino et al. 2015).

On the contrary, detecting insertions in the resequenced samples with respect to the reference genome is not straightforward. At the time we started this project, there were no dedicated tool allowing detecting new TE insertions. We therefore developed Jitterbug (Hénaff et al. 2015) and this program was used in the work here described. Jitterbug is a tool that identifies TE insertions in samples sequenced with paired-end approaches by selecting read pairs where one read maps to a unique genomic location in the reference genome, and the other maps at a discordant distance and it is similar to a TE found somewhere else in the reference genome (Figure 2.2). The results of TE-related polymorphism analyses are presented here. These studies provide more insights about the role of transposition in the recent evolution of these species' genomes.

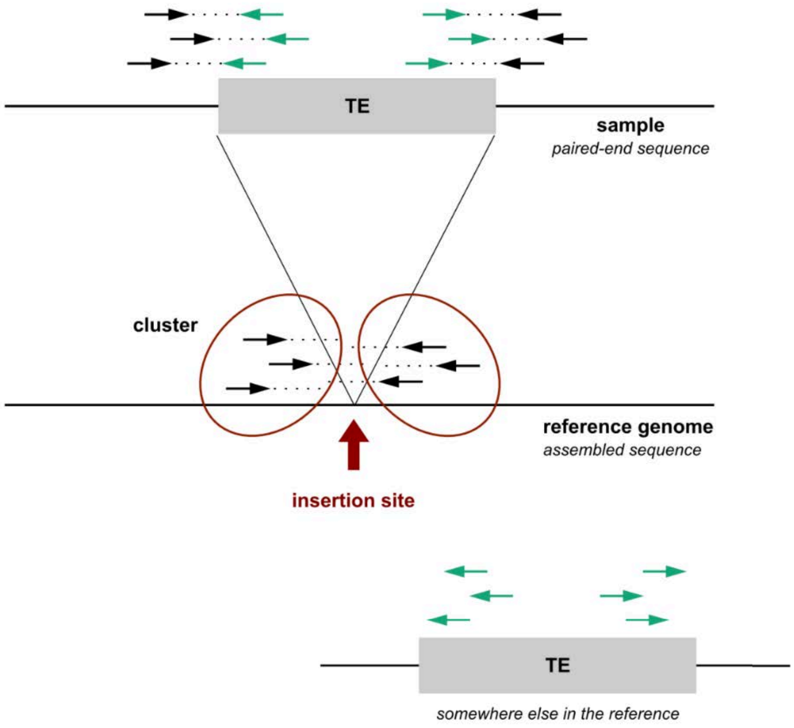


Figure 2.2. Discordant mapping read-pairs where one read maps to a unique genomic location (black) and the other maps to an annotated TE predict an insertion event (green), adapted from Hénaff et al. 2015.

2.4.1.- Melon

The melon reference genome is a double-haploid derived from the cross between PI 161375 (Songwhan charmi [SC]) (*conomon* group, ssp. *agrestis*) and the “Piel de sapo” line T111 (PS) (*inodorus* group, ssp. *melo*) (Garcia-Mas et al. 2012). In collaboration with the groups of Drs Jordi García-Mas and Sebastián Ramos-Onsins, we used resequencing data from seven melon accessions to analyze the TE-related species variability. In particular, our group was interested in determining the impact of TE insertions in the recent evolution of the melon genome.

The seven melon accessions analyzed come from different places and have evolved under different selective pressures (Table 2.1). The “Piel de sapo” line T111 (PS) and Védrentais (VED) may be considered as elite lines, the accessions Ames-24297 classified as *Cucumis trigonus* (TRI) and C-386 from Cabo Verde (CV) as wild lines, and accessions C-1012 from Irak (IRK), PI 161375 (Songwhan charmi [SC]) and PI 124112 (CAL) as landraces.

We have detected deletions (DEL) and insertions (INS) in the seven melon varieties (Table 2.4). As expected, the accessions PS and SC have been predicted less INS and DEL, around half compared to the other lines, because the reference genome is a double haploid line coming from a cross of these two lines.

Table 2.4. Number of insertions and deletions detected in the sequenced melon varieties

Accession	**Number of DEL**	**Number of INS**
CV	144	328
IRK	133	448
T111	108	288
SC	97	201
TRI	241	473
CAL	212	475
VED	187	475
TOTAL	**1122**	**2688**

Interestingly, there are more insertion predictions than deletions (Table 2.4), but as mentioned in the introduction, these terms are dependent on which genome is taken as reference or sample. Deletion detection may be more dependent on a high quality of the assembly, and the presence of miss-assembled repetitive regions or N islands could have a more important impact on the results. TEs should be properly annotated in the reference genome in order to detect the absence in the sample. However, insertions are probably less affected by unassembled regions, because they can only be detected in unique genomic regions and are more dependent on the coverage of the resequencing data.

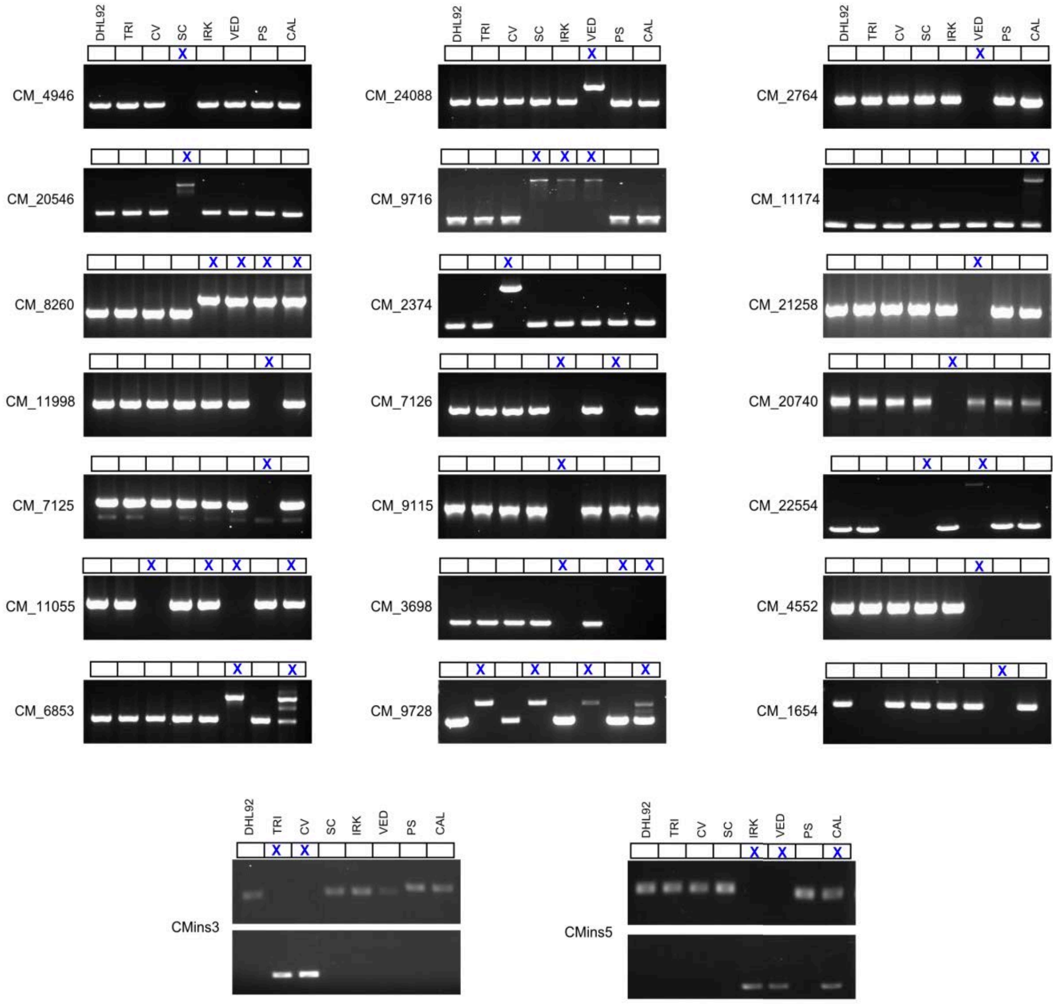


Figure 2.3. PCR validation of TE insertion predicted by Jitterbug. The name of the variety is indicated at the top, and the melon reference sample (DHL92) is included as a control. The TE insertion predicted by Jitterbug is indicated with a blue cross in the corresponding varieties. In case of CMins3 and CMins5, two PCRs were performed to reveal the empty (top) and the full (bottom) sites.

A total of 2,688 insertions were identified after running Jitterbug in the seven varieties, corresponding to 2,056 polymorphic loci. In order to confirm these results, 23 of the predicted polymorphic TE insertions were analyzed by PCR (Figure 2.3).

20 out of 23 were confirmed by PCR for the seven varieties. In some cases, double bands were amplified corresponding the presence and absence of the insertion in the same variety, suggesting that the insertion was not fixed in the variety (CM_20546, CM_8260, CM_6853, CM_24088, CM_9716, CM_2374, CM_9728 and CM_11174).

The predicted insertions and deletions from the seven melon varieties were combined into a set of polymorphisms. We defined a TE-related polymorphic locus (PM-TE) when either a deletion or an insertion at a particular locus has occurred in one or more lines with respect to the reference genome. We detected a total of 2,735 PM-TE the vast majority of which could be assigned to either retrotransposons or DNA transposons and only 2.8% of them couldn't be assigned due to their complex nature (Table 2.5).

Table 2.5. Classification of PM-TE into TE superfamilies

Superfamily	# PM-TE	% PM-TE	# copies in the genome	% copies in the genome
gypsy	661	24.17	28,174	23.68
copia	635	23.22	17,346	14.58
Non-LTR retrotransposon	10	0.37	129	0.11
retrotransposon fragment	469	17.15	42,733	35.91
Total of retrotransposons	**1,775**	**64.90**	**88,382**	**74.27**
CACTA*	93	3.40	6,944	5.84
hAT*	11	0.40	677	0.57
MULE*	239	8.74	9,497	7.98
Mariner*	3	0.11	609	0.51
Other MITEs	355	12.98	2,175	1.83
helitron	4	0.15	746	0.63
PIF*	147	5.37	3,094	2.60
DNA TE fragment	29	1.06	6,874	5.78
Total of DNA TE	**881**	**32.21**	**30,616**	**25.73**
uncategorized	**79**	**2.89**		
TOTAL	**2,735**	**100**	**118,998**	**100**

* including short elements that could be MITEs

Retrotransposons are at the origin of 64.9% of PM-TE consistent with the fact that retrotransposons are the most abundant TEs in melon genome. Both *Gypsy* (24.17%) and *Copia* (23.22%) contributed an important fraction of PM-TE, suggesting they have been active during the recent melon genome evolution. DNA transposons PM-TE account for the remaining 32.21%, being MITEs the most active ones.

Nine TE families are responsible for the 36.23% of the PM-TE, and these families represent less than 4% of the annotated TEs in the melon genome (Table 2.6).

Table 2.6. Most of the polymorphisms were caused by the mobilization of a small number of transposon families

Family	**Superfamily**	**PM-TE**	**%**	**Annotated**	**%**
CM_MITE_2617	CACTA (MITE)	224	8.19	700	0.59
CM_MULE_10	MULE	187	6.84	682	0.57
CM_gypsy_116	gypsy	120	4.39	177	0.15
CM_PIF_6	PIF	111	4.06	881	0.74
MELON_MITEs_1_43749	PIF (MITE)	110	4.02	117	0.1
CM_copia_96 *	copia	70	2.56	1695	1.42
CM_copia_45	copia	59	2.16	184	0.15
CM_copia_70 *	copia	59	2.16	39	0.03
CM_gypsy_137	gypsy	51	1.86	107	0.1
Total		**991**	**36.23**		**3.85**

* complex families composed of nested insertions

The most polymorphic families are 5 retrotransposon families, 2 MITE families and 2 DNA transposon families. When we compared the elements of each family by sequence similarity, we confirmed that these families are composed by relatively young elements. Interestingly, around 60% of PM-TE is present in only one variety (Table 2.7).

Table 2.7. Number of lines sharing TE polymorphic sites

	Number of lines sharing TE polymorphic sites					
	1	**2**	**3**	**4**	**5**	**6**
Retrotransposons	1,069	139	84	61	112	310
DNA transposons	561	103	41	29	34	113
uncategorized	0	37	29	12	1	0
TOTAL	1630	279	154	102	147	423

This is consistent with the data showing that most PM-TEs are due to young elements and confirms that TEs have been actively transposing during the recent evolution of melon, and probably during its domestication and the breeding of melon varieties.

In this study, we included varieties from two different melon subspecies, three from the ssp. *melo* and four from the ssp. *agrestis*. However, few PM-TE specific for each subspecies were detected (Table 2.8).

Table 2.8. TE polymorphic insertions located close or within genic regions

	Total PM-TE	**PM-TE < 500 bp**	**%**	**PM-TE in genes**	**%**	**PM-TE in exons**	**%**
All lines	2,735	826	30.20	611	22.34	361	13.20
agrestis* vs *melo	31	4	12.90	4	12.90	3	9.68
elite vs others	69	13	18.84	12	17.39	8	11.59
PS vs VED	671	231	34.43	165	24.59	105	15.65

Interestingly, there is an important fraction of PM-TE located in genic regions. Of the 2,735 PM-TE identified, 22% of these are found within genes and 7.8% are located within 500 nt of a gene (Table 2.8).

For instance, 165 PM-TE located in coding regions were identified between the two elite lines PS and VED (Table 2.8). The two elite lines analyzed, PS and VED, are closely related phylogenetically, but they differ in many agronomical traits, like fruit shape, flesh color, aromas, sugar content and ripening behavior (Sanseverino et al. 2015).

As a first step to investigate whether these TE insertions within genes and polymorphic between PS and VED may be associated with interesting traits, we analyzed the overrepresentation of GO categories in this dataset. GO enrichment analysis revealed that 165 genes, which are polymorphic due to a TE insertion between PS and VED, are related to sugar metabolism, hormone signaling and development of reproductive structures. In order to obtain more information about the function of these genes, we performed a tBLASTx of 165 genes against *Arabidopsis thaliana* genes (TAIR 9 gene annotation). We decided to focus on eight candidate genes potentially related to flower development (Table 2.9).

Table 2.9. List of genes impacted with a TE insertion polymorphic between VED and T111

	TE insertion in	**TE type**	**exon/intron region**	**ID gene**	**Description *C. melo***	**ID orthologous gene in *A. thaliana***	**Description orthologous gene in *A. thaliana***
CM_cd1	T111, VED	copia	intron	MELO3C003935	Similar to Superkiller viralicidic activity 2-like 2 (RNA helicase)	AT2G06990	Encodes a putative DExH-box RNA helicase that acts redundantly with HEN1, HUA1, and HUA2 in the specification of floral organ identity in the third whorl.
CM_cd2	VED	copia	intron	MELO3C018601	Similar to MADS-box protein SVP (Arabidopsis thaliana) (uniprot_sprot:sp|Q9FVC1|SVP_ARATH)	AT2G22540	Encodes a nuclear protein that acts as a floral repressor and that functions within the thermosensory pathway. SVP represses FT expression via direct binding to the vCArG III motif in the FT promoter.
CM_cd3	VED	copia	exon-3prime	MELO3C010848	Similar to Transcription factor HEC1 (Arabidopsis thaliana) (uniprot_sprot:sp|Q9FHA7|HEC1_ARATH)	AT5G67060	HECATE 1 (HEC1); FUNCTIONS IN: sequence-specific DNA binding transcription factor activity; INVOLVED IN: transmitting tissue development, carpel formation, regulation of transcription; LOCATED IN: nucleus
CM_cd4	T111	copia	intron	MELO3C011570	Similar to Alpha-xylosidase (Arabidopsis thaliana) (uniprot_sprot:sp|Q9S7Y7|XYL1_ARATH)	AT1G68560	Encodes a bifunctional alpha-l-arabinofuranosidase/beta-d-xylosidase that belongs to family 3 of glycoside hydrolases.
CM_cd5	VED	gypsy	exon	MELO3C008068	Similar to Patellin-3 (Arabidopsis thaliana) (uniprot_sprot:sp|Q56Z59|PATL3_ARATH)	AT1G72160	Sec14p-like phosphatidylinositol transfer family protein; FUNCTIONS IN: transporter activity; INVOLVED IN: transport; LOCATED IN: plasma membrane
CM_cd6	VED	copia	intron	MELO3C003188	Similar to Probable glutathione S-transferase (Glycine max) (uniprot_sprot:sp|P32110|GSTX6_SOYBN)	AT1G58602	LRR and NB-ARC domains-containing disease resistance protein; FUNCTIONS IN: ATP binding; INVOLVED IN: defense response, apoptosis
CM_cd7	VED	copia	intron	MELO3C006833	Similar to Predicted protein	AT5G51590	AT hook motif DNA-binding family protein; FUNCTIONS IN: DNA binding; INVOLVED IN: biological_process unknown; LOCATED IN: cellular_component unknown
CM_cd8	VED	MULE	exon-3prime	MELO3C024393	Similar to Transmembrane 9 superfamily member 4 (Mus musculus) (uniprot_sprot:sp|Q8BH24|TM9S4_MOUSE)	AT3G13772	Encodes an Arabidopsis Transmembrane nine (TMN) protein. Transmembrane nine (TM9) proteins are localized in the secretory pathway of eukaryotic cells and are involved in cell adhesion and phagocytosis. Overexpression of this protein in yeast alters copper and zinc homeostasis.

These eight PM-TEs were analyzed by PCR. The band for presence of PM-TE was amplified in 4 cases (CMcd3, CMcd4, CMcd7 and CMcd8), confirming the insertion of the TE within the gene. In the other 4 cases, we did not amplify the expected band. As it is unknown the length of the inserted TE, maybe the extension time of the PCR was not properly adjusted. Maybe we failed to amplify a unique product due to the repetitiveness of the analyzed loci. Also, different DNAs were used to perform the sequencing and PCR analysis, indicating the lack of fixation in the population.

We decided to investigate the expression during flower development of the genes with confirmed insertions. Both VED and PS are monoecious melon varieties and in the same plant there are feminine and masculine flowers. We collected flowers of both sexes at three different development stages and from different plants (Figure 2.4).

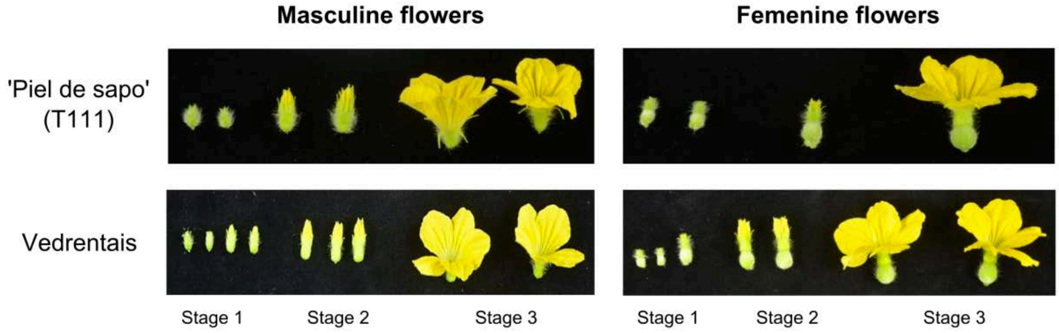


Figure 2.4. Developmental stages of the flowers from 'Piel de sapo' and 'Vedrentais' collected to analyze the expression of candidate genes. The flowers were picked at flower bud stage (Stage 1), intermediate stage bud-flower (Stage 2) and open flower (Stage 3).

To assess whether the presence of a TE affect the expression of the gene, RT-PCR analysis was performed for the four candidate genes. We failed to amplify a unique product of CMcd3, so we discarded this candidate gene. Of three analyzed cases, there's no significant difference between the gene expression of VED and PS (Figure 2.5), where an inserted TE is present in VED and absent in PS. The expression levels obtained in the candidate genes don't correlated with the presence or absence of the TE in the three flower stages from these two elite lines.

For two candidate genes (CMcd4 and CMcd7), the PM-TE is located in the intronic region, which may be spliced without affecting the expression levels. But the PM-TE of CMcd8 candidate gene overlaps with the coding sequence. Although this candidate gene can still be transcribed (Figure 2.5c), the TE insertion may affect the functionality

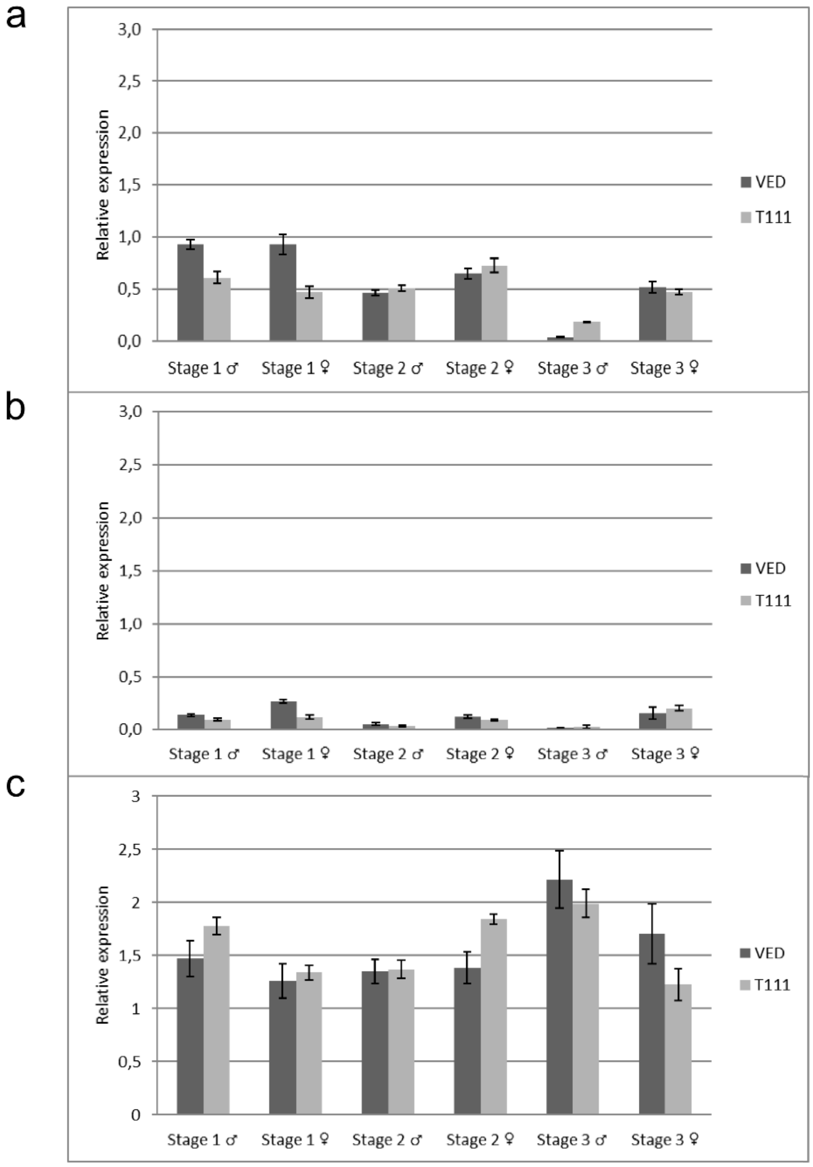


Figure 2.5. Expression of candidate genes in the different floral tissues of both elite lines (VED and T111). The candidate genes correspond to CMcd4 (a), CMcd7 (b) and CMcd8 (c), where in all cases a TE is present in VED and absent in T111. Relative transcript levels are given, following normalization with CYP7 values.

of the protein. More analysis should be performed to determine the impact of TE insertion in these candidate genes.

These results obtained suggest that the TE insertions, located within its candidate gene and polymorphic between PS and VED, did not altered the expression of the gene in the tested conditions. However, this information is a great source for studying the genetic variability due to TE insertion polymorphisms and for detecting genomic regions involved in domestication.

2.4.2.- Date palm

The group of Dr. Khaled Michel Hazzouri and Dr. Michael Purugganan (NYU, Abu Dhabi, United Arab Emirates) who are working on the evolution of date palm (*Phoenix dactylifera*) contacted us identify and analyze polymorphisms due to TE insertions and deletions in 69 varieties of date palm.

Date palm is a cultivated woody plant species and the first cultivation evidence was recorded in 3,700 BC in the area between the Nile Rivers and Euphrates (Al-Mssallem et al. 2013). Date palm was introduced by humans to northern India, north Africa and southern Spain. The 69 varieties of this analysis come from different regions of Arabic countries and some of them can be considered as landraces or domesticated lines. Among them, there is a wide range of phenotypes, of which the most important agronomical traits are date quality and yield (Al-Mssallem et al. 2013).

The date palm reference genome corresponds to elite ‘Khalas*’* cultivar (Al-Mssallem et al. 2013). The assembled genome size is 558Mb and covers around 83% of the estimated genome size (around 671 Mb), which contains 82,354 scaffolds (N50 = 329.9kb) and contigs less than 500 bp were discarded. The annotation (Hazzouri, unpublished) consists of 25,059 predicted genes and a total of 274,357 TE copies, where the 80% of them corresponds to retrotransposons (Table 2.10).

Table 2.10. Classification of transposable elements in date palm reference genome

	# copies in the genome	**% copies in the genome**
Total retrotrtansposons	**221,521**	**80.74**
LTR	175,612	64.01
Copia	97,448	35.52
Gypsy	73,326	26.73
TRIM_LARD	4,838	1.76
noLTR	45,909	16.73
LINE	38,937	14.19
SINE	6,972	2.54
Total DNA transposons	**32,922**	**12.00**
CACTA	3,611	1.32
hAT	11,920	4.34
MuDR	1,905	0.69
Tase	5,777	2.11
MITE	4,968	1.81
Helitron	3,930	1.43
uncategorized	**19,914**	**7.26**
TOTAL	**274,357**	**100**

A total of 69 date palm varieties were used to detect insertions and deletions respect to the reference genome. The 69 varieties were resequenced by paired-end Illumina sequencing. The 69 *bam* files were provided by Dr. Hazzouri, and each of them had a very different coverage, ranging from 6x to 58x (Table 2.11).

Table 2.11. List of analyzed varieties, indicating the coverage of *bam* files

Samples	**Coverage bam file (x)**	**StDev bam file**
Hayany	58.30	129.30
Amir_haj	51.68	114.71
Samany	44.81	102.35
Zahidi	43.91	109.31
Abouman	39.91	115.87
Mazafati	38.62	70.26
Tagiat	35.19	62.50
Began	32.04	74.43
Khisab	31.60	116.67
Halawy	31.17	98.91
Nebeit seif	28.91	80.87
Khadrawy	28.76	75.70
Biddajaj	28.24	71.89
Karbali	28.05	58.68
Fard4	27.08	68.40
Khenezi	26.94	98.92
Faslee	26.85	61.60
Ruth	26.46	66.68
Metasealth	26.17	85.57
Zagloul	26.16	73.12
Sultana	25.36	75.28
Khastawi	25.30	86.33
Judah	25.22	63.10
Rhars	25.18	73.70
Aseel	24.80	62.37
Jonah	24.70	66.94
Kuproo	24.47	66.79
Jeremiah	24.41	61.43
Besser haloo	24.09	100.94
Rothan	24.01	72.32
Jao	23.97	54.93
Braim	23.90	75.15
Lulu	23.83	96.99
Horra	23.77	80.85
Abel	23.55	64.53
Boaz	23.51	69.95
Eve	23.10	56.37
Khasoy	23.08	54.77
Thory	23.03	64.51
Aziza	22.20	50.69
Saidi	22.16	74.76
Alig	22.10	77.15
Kashoowari	22.06	70.25
Deglet noor	21.63	58.15
Ajwa	21.27	69.23
Naquel khuh	20.86	64.86
Shagri	19.98	61.83
Barhee	19.54	52.71
Medjool	19.26	61.43
Fagous	19.02	46.97
Um al hamam	17.79	55.35
Silani	17.79	50.62
Piavom	17.00	53.41
Dayri	16.97	54.80
Chichi	16.17	55.00
Nagal	15.44	48.39
Maktoumi	15.28	62.95
Helwa	14.95	54.77
Rabee	13.66	36.24
Dedhi	13.42	38.09
Ebrahimi	13.22	34.21
Um al blaliz	13.03	33.37
Azraq azraq	12.72	33.97
Ewent ayob	12.12	33.44
Dajwani	11.71	31.87
Dibbas	10.85	49.62
Kabkab	9.22	29.48
Hilali	7.00	32.66
Hiri	6.01	25.62

We have detected insertions and deletions in the 69 date palm varieties. While Pindel software seems to be highly dependent on coverage to detect deletions, Jitterbug can accurately predict TE insertion given different coverage (Figure 2.6). This difference is mainly due to the quality of the reference genome. In case of deletions, TEs should be properly annotated in the reference genome in order to predict absence in the sample. The unassembled regions or N islands interfere with the proper TE detection and annotation and consequently also with the prediction of deletions in the sample. However, insertions are not that much affected by unassembled regions, because Jitterbug predicts insertions in unique genomic regions which are usually well assembled.

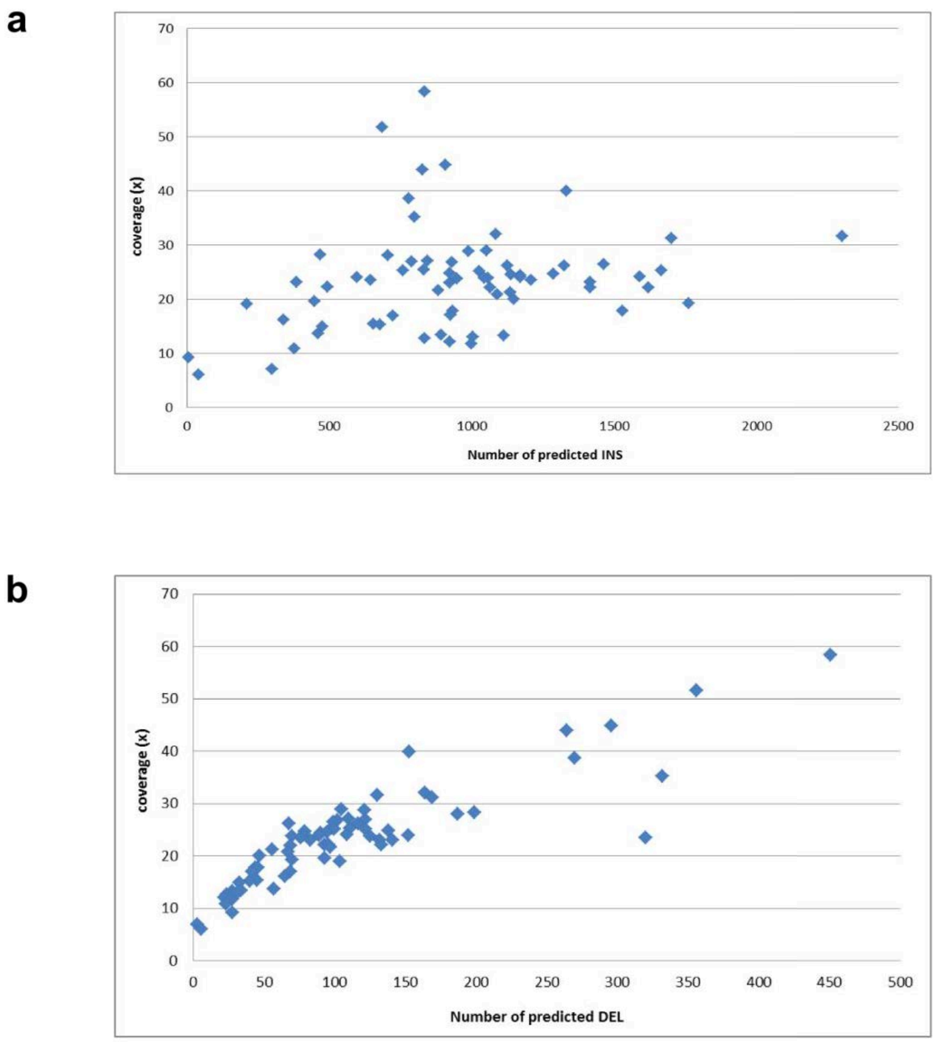


Figure 2.6. Correlation between the coverage of bam files versus the number of predicted insertions (a) and deletions (b).

A total of 117,435 insertions and 7,616 deletions were detected in the 69 date palm varieties (Figure 2.7). More than 92% of insertions are found in the 1000 longest scaffolds, except for ‘Hilali’, ‘Hiri’ and ‘Kabkab’, which have the lowest coverages. The number of insertions varies greatly between lines, with the ‘Khisab’ variety having 3,629 insertions and the ‘Kabkab’ variety only 7 insertions. And same diversity between lines is found in deletions, with the ‘Hayany’ variety having 451 deletions and the ‘Hilali’ variety only 3.

Both deletions and insertions can be classified according to the superfamilies the inserted/deleted TE belongs to. The vast majority of insertions and deletions are classified as *Gypsy* and *Copia* LTR retrotransposons in all varieties (between 50% and 60%), except for the three varieties with the lowest coverage (Figure 2.8 and 2.9). The total number of insertions and deletions in these three varieties were too low to allow a proper analysis.

Around 20% of insertions and deletions in all varieties are classified as LINEs, consistent with the fact that LINEs are the most abundant TEs, after *Gypsy* and *Copia* elements, in the date palm genome. Amongst DNA transposons, MITEs are more highly represented than their relative proportion of annotated copies in the genome (1.81%), indicating that are quite active.

Apart from the interest to study the TE impact in date palm genome, this analysis allowed us to study the influence of the coverage on the predictions of insertions and deletions. From these results, a coverage of 25x could be considered good enough for a proper analysis of TE insertion polymorphisms.

Currently, this data is being used to study the possible TE impact on genes responsible for important agronomical traits in date palm in the Dr. Purugganan’s lab.

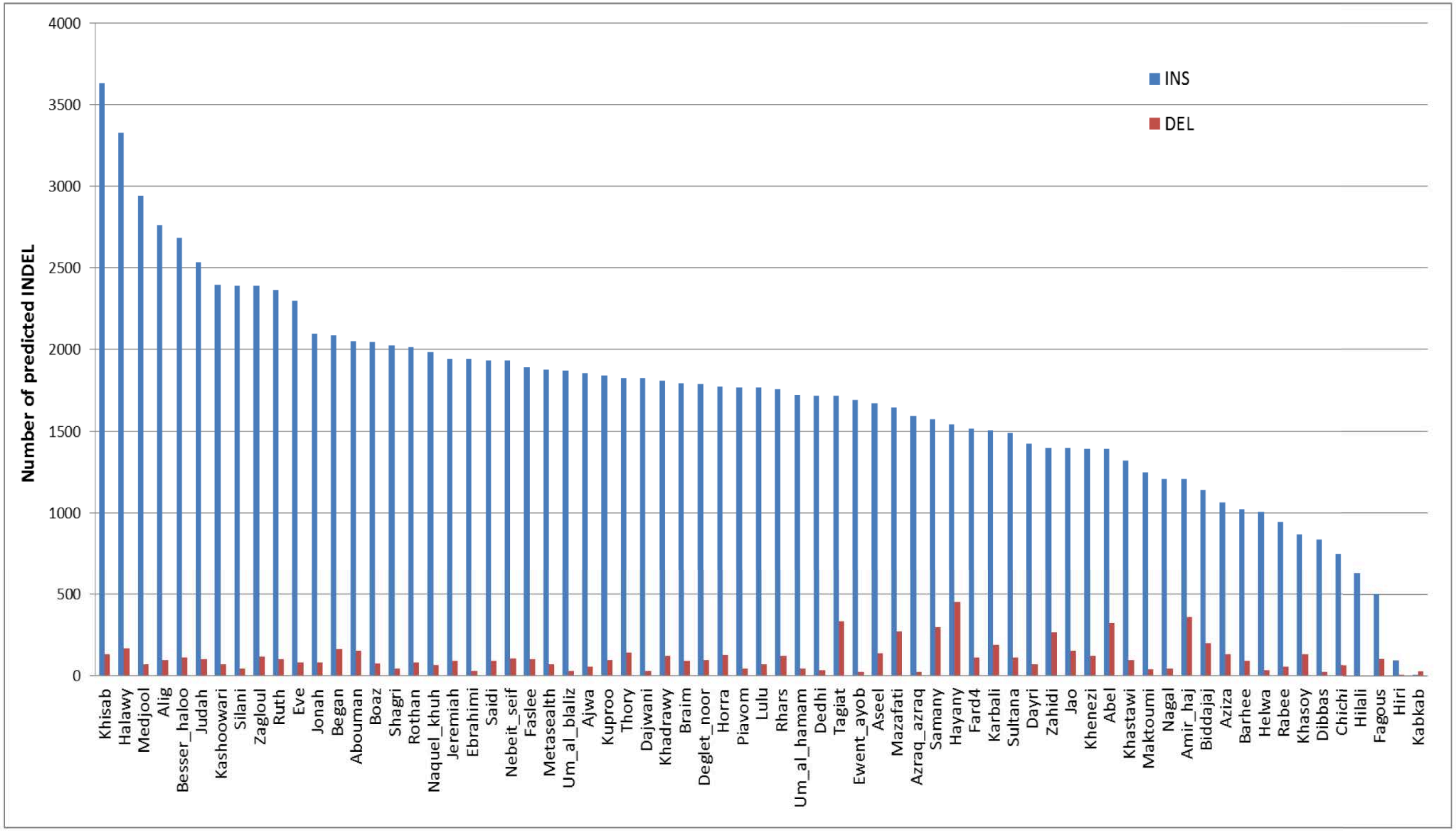


Figure 2.7. Number of insertions and deletions

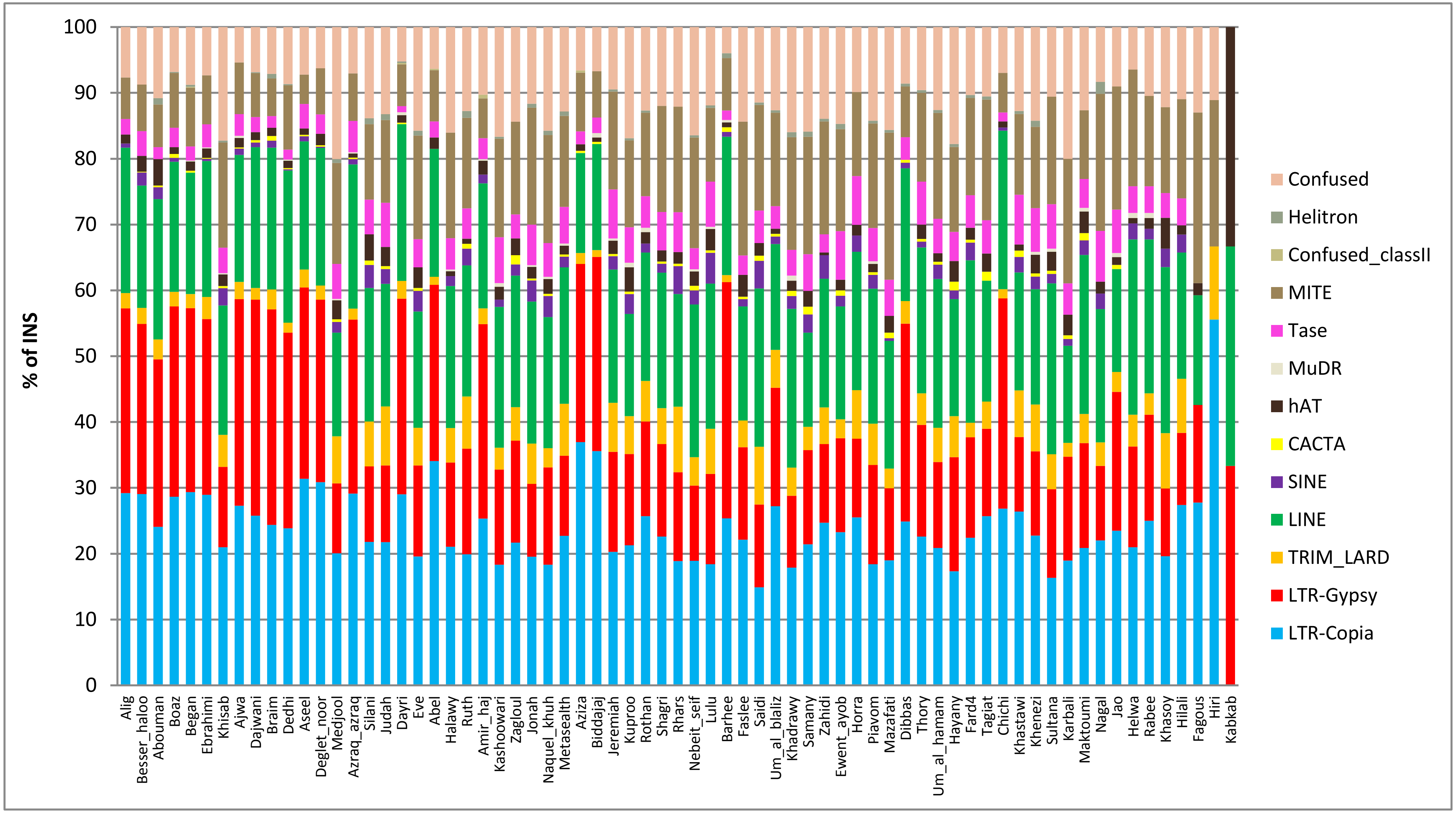


Figure 2.8. Classification of insertions into superfamilies

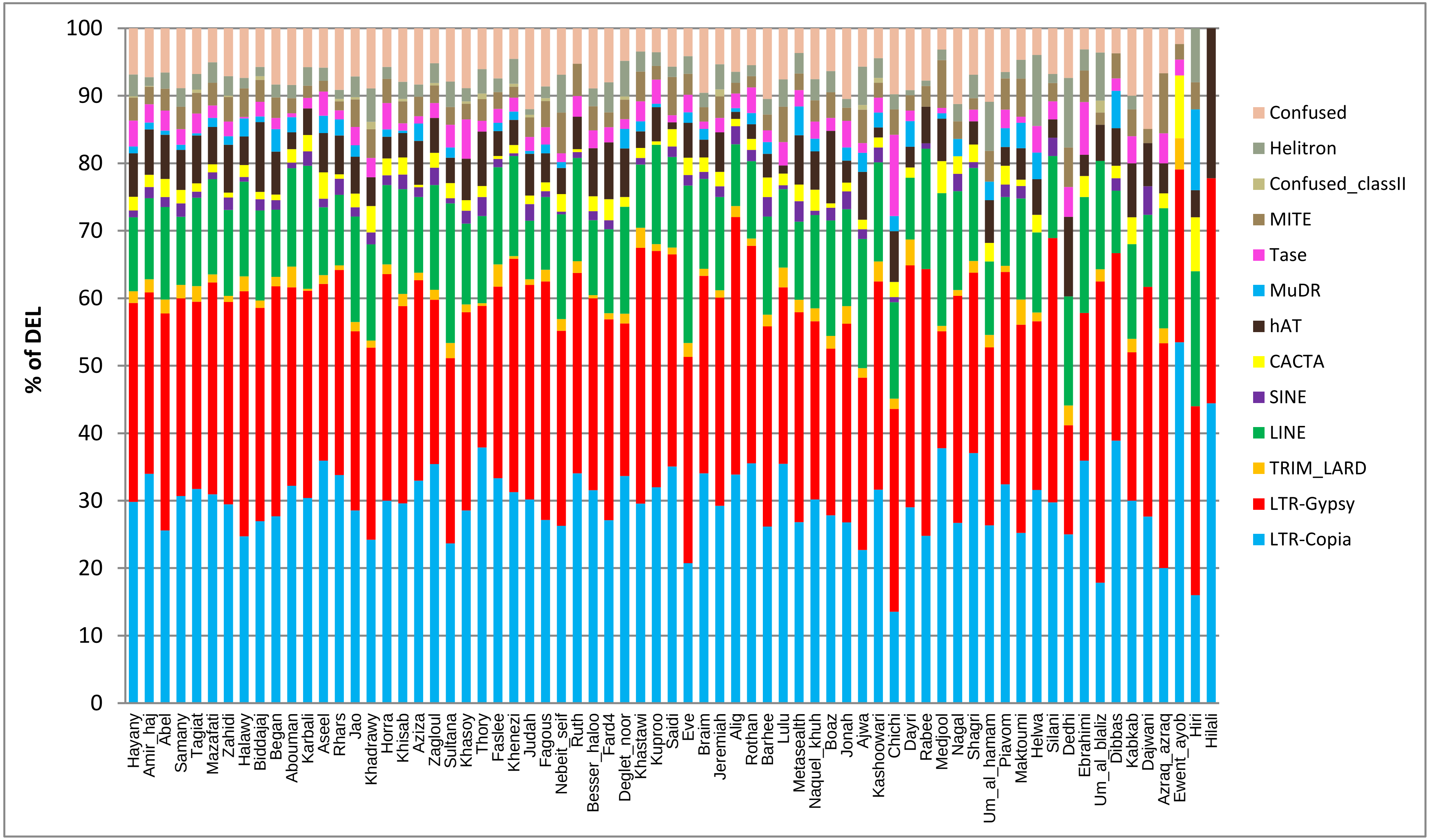


Figure 2.9. Classification of deletions into superfamilies

2.4.3.- *Physcomitrella patens*

An international consortium led by Dr. Stefan A. Rensing (University of Marburg, Germany) has obtained a new version of *Physcomitrella patens* reference genome (Rensing et al. submitted). Our group has been in charge of different analysis on the dynamics of TEs in the moss genome, including the analysis of polymorphisms due to transposons.

The moss *Physcomitrella patens* is a non-vascular multicellular land plant and a member of the bryophyte family which diverged from the land plant lineage more than 400 Mya (Kenrick and Crane 1997). This species has become a model system to study plant development, growth and cell differentiation (Sakakibara et al. 2003; Repp et al. 2004).

P. patens is a dominant haploid species and its genome is composed by 27 chromosomes. The assembled genome size is 462.3 Mb, totaling 89% of the 518 Mb estimated by flow cytometry (Schween et al. 2003). The reference genome corresponds to *P. patens* ssp. *patens* 'Gransden 2004' strain, which is commonly used in many laboratories (Ashton and Cove 1977). There are 35,938 predicted genes which occupy the 16.9% of the genome space.

For this new genome version, two TE annotations have been obtained using different approaches. On the one hand, the Main TE annotation (data from Heidrun Gundlach, PGSB Germany) has identified structural TE features in order to discriminate between TEs and non-TE repeated sequences, being a more accurate annotation. On the other hand, the REPET annotation (data from Dr. Florian Maumus, URGI-INRA France) combines structural features and similarity searches to annotate TEs and non-TE repeated sequences, being a more comprehensive annotation. The TE fraction corresponds to 54.22% or 52.16% of the genome, depending whether using the Main TE annotation or the REPET TE annotation. The vast majority of annotated TEs in both annotations correspond to LTR retrotransposon *Gypsy*-like elements (Table 2.12).

Table 2.12. Classification of transposable elements obtained using two different approaches in *P. patens* reference genome.

		Main TE annot		**REPET TE annot**	
	Code	**# annotated TEs**	**% of annotated TEs**	**# annotated TEs**	**% of annotated TEs**
Class I		**214,553**	**88.73**	**289,844**	**93.78**
LTR Copia	RLC	17,375	7.19	16,966	5.49
LTR Gypsy	RLG	150,776	62.36	267,235	86.46
Unclassified LTR	RLX	41,306	17.08	33	0.01
LINE	RIX	546	0.23	363	0.12
SINE	RSX	8	0.003		
unclassified retrotransposon	RXX	4,542	1.88	5,247	1.70
Class II		**20,094**	**8.31**	**4,814**	**1.56**
Helitron	DHH	2,789	1.15	1,575	0.51
Harbinger-PIF	DTH	10,647	4.40		
Unclassified DNA-TIR transposon	DTX	756	0.31	2,309	0.75
Unclassified DNA transposon	DXX	5,902	2.44	919	0.30
Unclassified Transposable element	**TXX**	**7,154**	**2.96**	**14,421**	**4.67**

Transposon insertions and deletions were analyzed comparing the resequenced variety ‘Villersexel’ with respect to the reference genome variety ‘Gransden’ (Rensing et al. submitted). The paired-end reads of *P. patens* accession ‘Villersexel’ were downloaded from the SRA database of NCBI with SRP004339 as project code, which contains two experiments and a total of 5 runs (Table 2.2).

We treated the 5 runs of resequencing variety as independent samples generating a *bam* file per each one and Jitterbug was run separately using the Main TE annotation as a first approach (Table 2.13). We compared runs from the same experiment in order to know whether predicted insertions were the same. In case of experiment SRX037761, a total of

Table 2.13. Number of TE-INS detected by Jitterbug per each run bam file using the Main TE annotation.

Experiment Code	Run Code	Coverage bam file (x)	StDev bam file	# TE-INS
SRX037761	SRR090654	4	7.9	133
	SRR512764	16.08	31.31	270
SRX030894	SRR072296	8.62	16.72	236
	SRR191864	15.27	28.87	255
	SRR400524	8.35	15.92	216

292 insertions were predicted between the 2 runs, sharing 38% of predicted TE insertions (Figure 2.10a). Considering that one run has a lower average of coverage (SRR090654), fewer TE insertions (133) were predicted and almost all of them (83% of 270 predicted TE insertions) were found in the other run (SRR512764). In the second experiment SRX030894, 168 out of 300 predicted insertions are shared among the 3 runs (56%) (Figure 2.10b). Both run SRR072296 and run SRR400524 have the same coverage from *bam* files, and almost all insertions from these 2 runs (236 and 216, respectively) are predicted in the run with higher coverage from *bam* file (SRR191864, 255 predicted TE insertions).

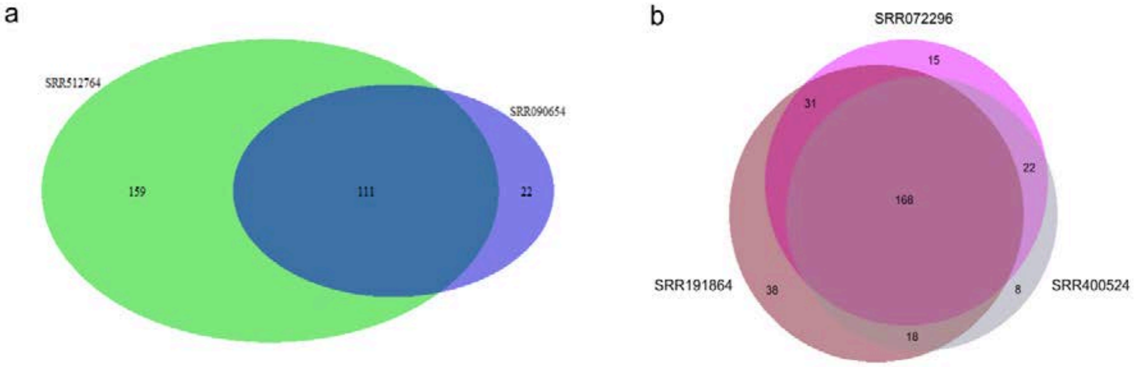


Figure 2.10. Transposon insertions shared when comparing runs from the same experiment using the Main TE annotation. The SRX037761 experiment consist on 2 runs that share 38% of TE-INS (a), and SRX030894 experiment is composed for 3 runs that share 56% of TE-INS predicted (b).

So, runs with high coverage contain most predicted transposon insertions and to simplify the analysis, we created a *bam* file per each experiment merging the corresponding *bam* files, which resulted in higher coverage. Also, we run Jitterbug using the Main TE annotation for both experiments separately. 227 out of 426 predicted TE insertions (65%) were shared by 2 experiments (Figure 2.11). This gave us the idea that *bam* files from each experiment are not saturated. We considered each experiment as an independent sample, because it's not well specified in SRA database what kind of tissue has been used.

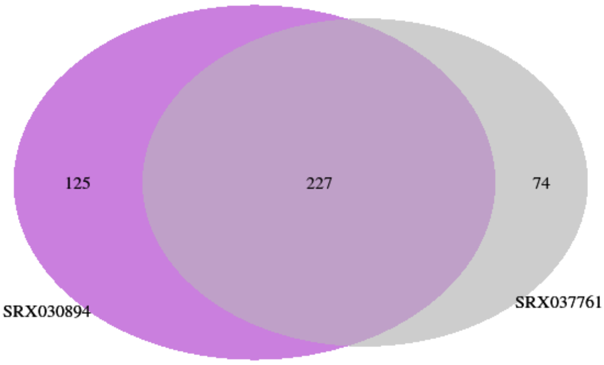


Figure 2.11. Transposon insertion predicted using the Main TE annotation and a *bam* file per each experiment.

In order to analyze the impact of the TE annotation on the prediction of TE polymorphisms we took advantage of the fact that our group have access to two *P. patens* TE annotations obtained with different approaches: the Main TE annotation and the REPET annotation.

We run Jitterbug and Pindel, respectively, using *bam* files created per each experiment and for both TE annotations. The comparison of the results obtained showed that much more deletions and insertions were predicted using the REPET TE annotation with respect to using the Main TE annotation (between 20 and 28% more) in both experiments (Table 2.14).

Table 2.14. Comparison of number of TE-INS and TE-DEL obtained in each experiment using different TE annotation.

			Main TE annotation		REPET TE annotation	
Experiment Code	Coverage bam file (x)	StDev bam file	# TE-INS	# TE-DEL	# TE-INS	# TE-DEL
SRX037761	19.33	35.67	301	1006	420	1396
SRX030894	31.73	47.71	352	1643	444	2298

In addition, more deletions have been predicted rather than insertions in both TE annotations. One possible explanation could be that whereas insertions cannot be detected in repetitive regions because of one read maps to a unique genomic location in the reference genome, deletions can be identified in these regions.

In order to consider as TE polymorphism, only predicted deletions overlapping with annotated TEs were kept. Around 50% of predicted TE deletions overlap with only one annotated transposon from the Main TE annotation, while using the REPET TE annotation is almost 40% of the predicted TE deletions. Calculating the percentage of genome which corresponds to transposons, the difference between TE annotations is about 2.6% (being 54.22% TE occupancy for the Main TE annotation and 52.16% for the REPET TE annotation). This suggests us that the REPET TE annotation consist of a much higher number of annotated TEs but small elements.

We compared the predicted TE insertions and deletions using two different TE annotations (Figure 2.12).

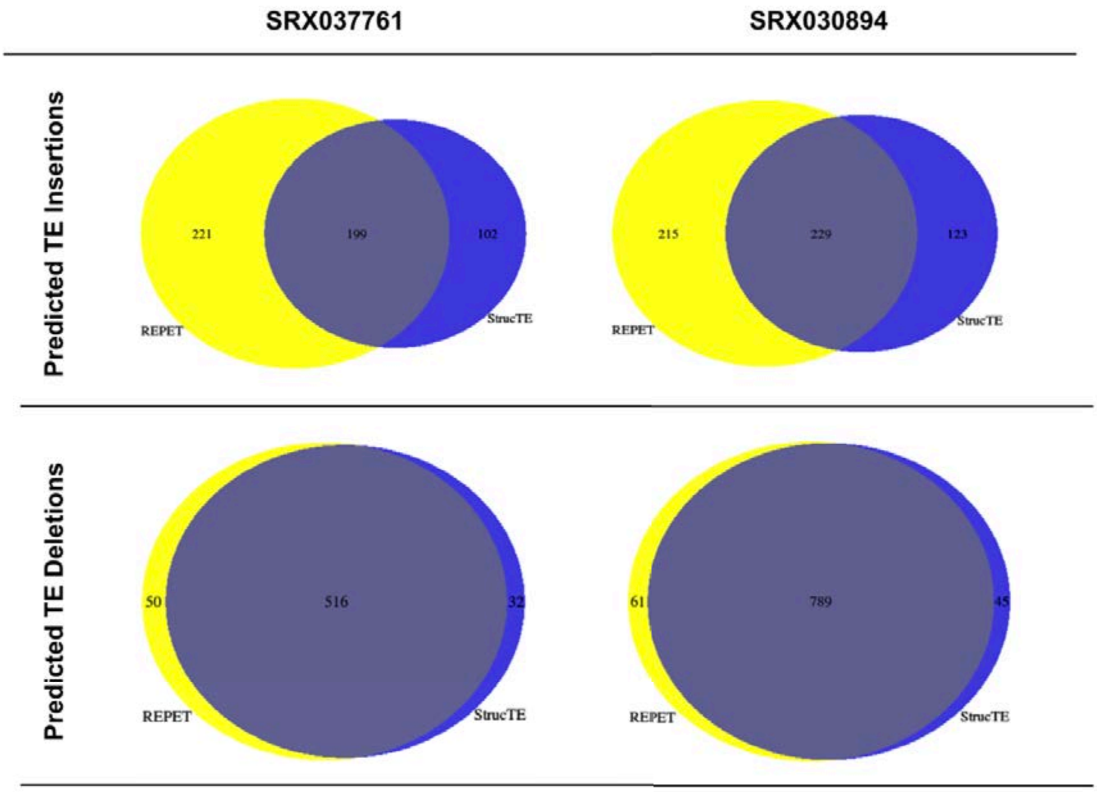


Figure 2.12. Comparison of TE annotation used for the predictions of TE polymorphism per each experiment. In yellow, REPET TE annotation and Main TE annotation in blue.

The vast majority of deletions were predicted in both cases (91% for SRX037761 experiment and 93% for SRX030894 experiment), while in case of predicted TE insertions only 47% for SRX037761 experiment and 52% for SRX030894 experiment were common for the two TE annotations. This is because the coverage may be too low and some predicted TE insertions may have few supporting reads in some samples and therefore they may have been missed.

Table 2.15 shows the classification in superfamilies of either insertions or deletions, in order to compare the 2 different TE annotations. In case of deletions, we have considered all TEs that overlap with one predicted deletion. The majority of insertions and deletions are classified as class I (between 91.48% and 96.32%). For the REPET TE annotation, there were no class II predicted insertions, and the major part of class I insertions and deletions correspond to *Gypsy* elements.

Table 2.15. Comparison of predicted insertions and deletions between the Main and the REPET TE annotation.

		SRX037761				**SRX030894**			
		Main TE annot		**REPET TE annot**		**Main TE annot**		**REPET TE annot**	
	Code	**INS (%)**	**DEL (%)**	**INS (%)**	**DEL (%)**	**INS (%)**	**DEL (%)**	**INS (%)**	**DEL (%)**
Class I		**290 (95.39)**	**873 (97.11)**	**410 (98.8)**	**975 (95.49)**	**337 (95.47)**	**1441 (97.3)**	**425 (98.15)**	**1661 (95.51)**
LTR *Copia*	RLC	13 (4.28)	52 (5.78)	12 (2.89)	45 (4.41)	6 (1.7)	97 (6.55)	12 (2.77)	99 (5.69)
LTR *Gypsy*	RLG	270 (88.82)	709 (78.87)	397 (95.66)	913 (89.42)	318 (90.08)	1137 (76.77)	412 (95.15)	1541 (88.61)
Unclassified LTR	RLX	7 (2.30)	100 (11.12)			13 (3.68)	193 (13.03)		
LINE	RIX		2 (0.22)		2 (0.2)		4 (0.27)		4 (0.23)
SINE	RSX								
Unclassified retrotransposon	RXX		10 (1.11)	1 (0.24)	15 (1.47)		10 (0.68)	1 (0.23)	17 (0.98)
Class II		**13 (4.28)**	**3 (0.33)**	**0**	**8 (0.78)**	**15 (4.25)**	**6 (0.41)**	**0**	**6 (0.35)**
Helitron	DHH	10 (3.29)	2 (0.22)		7 (0.69)	12 (3.4)	3 (0.2)		5 (0.29)
Harbinger-PIF	DTH								
Unclassified DNA-TIR transposon	DTX	1 (0.33)				1 (0.28)			
Unclassified DNA transposon	DXX	2 (0.66)	1 (0.11)		1 (0.1)	2 (0.57)	3 (0.2)		1 (0.06)
Unclassified Transposable element	**TXX**	1 (0.33)	23 (2.56)	5 (1.2)	38 (3.72)	1 (0.28)	34 (2.3)	8 (1.85)	72 (4.14)
	Total	**304**	**899**	**415**	**1021**	**353**	**1481**	**433**	**1739**

To obtain a set of transposon insertion polymorphism (PM-TE), we combined the insertion and deletion predictions of the two experiment samples. We identified a total of 1,390 PM-TE using the Main TE annotation and 1,572 PM-TE using the REPET TE annotation (Table 2.16).

Table 2.16. Number of total insertions and deletions due to a TE and also, it has been counted the number of PM-TE at different distances from the closest gene and its percentage

				Number of PM-TE at a certain distance of an annotated gene					
	# PM-INS	# PM-DEL	# PM-TE	more than 1kb	%	less than 1kb	%	within genes	%
Main TE annot	426	964	1390	1241	89.28	150	10.79	37	2.66
REPET TE annot	589	983	1572	1411	89.76	162	10.31	42	2.67

An important fraction of the TE-related polymorphisms belongs to the five families of LTR retrotransposons identified in the Main TE annotation (Table 2.17). The most polymorphic seems to be the *Gypsy* family called RLG1. Most of RLG1 elements are extremely recent and account for a quarter of the genome.

Table 2.17. Fraction of polymorphic TE insertions classified per TE type and family which have been identified in the Main TE annotation.

LTR retrotransposon	Family code	% of genome	% of annotated TEs	Num. PM-TE	% PM-TE
Gypsy	RLG1	25.49	47.11	602	48.51
	RLG2	5.56	10.27	72	5.80
	RLG3	9.16	16.94	72	5.80
Copia	RLC4	0.68	1.26	7	0.56
	RLC5	1.94	3.58	31	2.50
Subtotal		42.83	79.16	784	63.17

More or less 10% of PM-TEs are found less than 1kb from a gene, which 37 and 42 PM-TEs are located in coding regions, using the Main TE annotation and the REPET TE annotation, respectively. So, we compared whether the PM-TE inside genes are shared for both TE annotations. Only 23 PM-TE inside a gene were detected by both annotations, in which some of them have been predicted their functionality (Table 2.18). The presence-absence variations may potentially affect the coding capacity or the expression profile of these genes and possibly cause some phenotypic changes between both moss accessions.

Table 2.18. List of genes with a PM-TE in coding regions found in both TE annotations. For each gene, there are ID code of *P. patens* gene, ID code of orthologous gene in *A. thaliana* and the description of the gene from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ppatens)

ID_Ppatens	**ID_Ath**	**Description**
Phpat.002G012700	AT3G02750	Protein phosphatase 2C family protein
Phpat.002G031900	AT3G26750	HECT-like Ubiquitin-conjugating enzyme (E2)-binding
Phpat.002G032000	AT4G33380	
Phpat.002G114500	AT4G26690	PLC-like phosphodiesterase family protein
Phpat.002G119500		mRNA capping enzyme, catalytic domain
Phpat.003G043700		
Phpat.003G101900	AT1G03160	MSS1/TRME-RELATED GTP-BINDING PROTEIN
Phpat.004G021500		
Phpat.004G040700	AT2G14720	vacuolar sorting receptor 4
Phpat.004G094500	AT1G15360	Integrase-type DNA-binding superfamily protein
Phpat.004G115600	AT5G23440	ferredoxin/thioredoxin reductase subunit A (variable subunit) 1
Phpat.006G015200		
Phpat.006G015300	AT4G10465	Heavy metal transport/detoxification superfamily protein
Phpat.008G035000	AT1G63700	Protein kinase superfamily protein
Phpat.009G058700	AT4G16340	guanyl-nucleotide exchange factors;GTPase binding;GTP binding
Phpat.016G039900		
Phpat.016G077100	AT5G20920	eukaryotic translation initiation factor 2 beta subunit
Phpat.018G046700		
Phpat.019G038300	AT1G07020	
Phpat.021G018500	ATCG00700	photosystem II reaction center protein N
Phpat.022G075900		
Phpat.024G055700		
Phpat.026G038000	AT3G11170	fatty acid desaturase 7

2.5.- DISCUSSION

Due to their mobility, TEs generate insertion polymorphisms which are an important source of genetic variability that can be used in evolution. Resequencing crop varieties have become a useful approach to analyze genetic variation. To detect new TE insertions using NGS data several computational methods have been developed in the last years (Ewing 2015). These methods detect polymorphisms that appear as insertions or deletions with respect to the reference genome. Whereas detecting deletions with respect to the reference is relatively straightforward, detecting insertions is more complicated. Most software tools to detect insertions are adapted to analyze short-read paired-end resequencing data generated on Illumina platforms or similar. These programs take advantage to discordant mapped paired-end reads in which one of the read pairs maps to a unique genomic a location near the insertion site and the other maps to an annotated TE located elsewhere in the reference genome. All these bioinformatic programs can only detect insertions in non-repetitive regions. As the insertion predictions rely on a well-mapped pair while the other can map in many different places and elsewhere in the genome, insertions near an N island, nested TE insertions or low-complexity regions will generate conflicting results. It is important to take this limitation into account when discussing the genome distribution of polymorphic sites.

Jitterbug is a pipeline developed in our group that predicts novel TE insertions in a sample compared to reference genome. Jitterbug only requires a *bam* file and a GFF TE annotation file of the reference genome, whereas other programs require specific formatting files and/or need distinct input files (Hénaff et al. 2015; Rishishwar et al. 2016). We have used the Jitterbug program to detect insertions, together with the Pindel program (Ye et al. 2009) to detect deletions with respect to the reference genome. The use of these two programs has allowed us to analyze TE insertion polymorphisms in three different species, having different quality of genome assembly, TE annotation and resequencing data.

Repetitive genomic regions are a major problem in genome assembly and most plant genomes are fragmented in contigs or scaffolds with a lot of gaps. We have performed the analysis in three different genomes, ranging from the highly-fragmented date palm genome (N50=329.9kb) to a good-quality melon genome (N50= 4680kb) (Table 2.19).

Table 2.19. Characteristics of the plant reference genomes used in this chapter.

Common name	**Scientific name**	**Phyla**	**Size (Mb)**	**Scaffold N50 (kb)**	**Gene (#)**	**TEs (%)**	**Reference**
Melon	*Cucumis melo*	Dicot	450	4,680	27,427	19	Garcia-Mas et al. 2012
Date palm	*Phoenix dactylifera*	Monocot	671	329.9	25,059	n.d.	Al-Mssallem et al. 2013
Moss	*Physcomitrella patens*	Bryophyte	510	1,320	35,938	58.25	Rensing et al. submitted

The genome quality affects the proper TE detection and annotation. As a first step, a library of consensus sequences is generated and is representative of TEs found in the genome. The more fragmented is the genome, the more difficult it is to create a consensus database and to annotate the copies in the genome. TE content will consist on several partial annotations in highly-fragmented genomes instead of a single complete one. This can have some implications in performing a proper TE-related polymorphism analysis. For instance, in order to consider as TE polymorphism, only predicted deletions overlapping with annotated TEs were kept. Deletions with more than one TE annotation may correspond to partially annotated TE, but also to deletions containing TEs but not generated by a TE movement. These facts should be taken into account in detecting and annotating TEs in highly fragmented genomes.

It’s well known that TEs represent an important fraction of plant genomes. But TEs are annotated using different methods and parameters, and therefore the annotations of different genomes are not directly comparable. We have used two TE annotations obtained from different approaches in *P. patens* analysis. A more detailed but limited annotation that we named “Main TE annotation” and a more thorough annotation obtained using the REPET pipeline. Despite using different annotations, we have found more or less the same number of TE-related polymorphisms (1,390 Main TE annotation vs 1,572 REPET annotation). While the “Main TE annotation” have allowed us to classify the TE-related polymorphisms and to study particular families or elements, the REPET annotation is composed of smaller annotation, making difficult to study particular families. However, the REPET annotation can be used for masking the genome in order to annotate genes or study the landscape, like in Chapter 1.

Another important aspect performing these analysis is the coverage of resequencing data used. The analysis in date palm allowed us to check how coverage affects TE insertion polymorphism detection. Besides certain resequencing samples, the presence of reads that support predictions was enough with 25-fold depth of coverage. But this threshold is very arbitrary, and an insertion cannot be detected due to the lack of reads for that particular region or unmappability in that region. In the *P. patens* analysis, we combined *bam* files in order to simplify the experiment and to have more coverage. The combination of *bam* files up to a coverage of 19.33x and 31.73x per each experiment did not result in a saturation of the insert identification.

All these analyses presented above show some challenges and allow us to determine the required conditions for a proper TE-related polymorphism analysis.
The quality of reference genome sequence affects directly to the proper TE annotation. The repetitive sequences are mainly the unassembled fraction, and TE fraction may be underestimated.

Linked to that, the correct mapping of resequencing reads on the reference genome sequence affects to the detection of TE polymorphisms. Jitterbug relies on insert size determined by the discordant reads and the split reads. It’s important to check whether *bam* files of pair-end reads are homogeneous in interval size, otherwise the number of false positive may increase.

One limitation to the detection of TE polymorphisms is that they are only detected in non-repetitive regions of the analyzed genome. Moreover, the ability to detect TE polymorphisms will depend on the diversity between the samples and the reference genome. To sum up, a proper analysis of TE polymorphism requires a good quality reference genome with its well annotated TE fraction and a *bam* file with a certain coverage of reads.

The analysis of transposon insertion polymorphisms has provided a rich amount of information about dynamics of transposon activity. Our results suggest that retrotransposons are the most polymorphic TEs in melon, date palm and moss genomes. This may be in part due to the fact that in all three analyzed genomes, retrotransposons are the most predominant type of TEs, as this is also the case in most plant genomes (Lisch 2012). But these results may also suggest that LTR retrotransposons have been particularly active during the recent evolution of these three crops.

Whereas the analysis of date palm was limited to the study of the influence of the coverage in the reliability of the PM-TE identification, our group has a more general interest in studying the recent evolution of melon and moss. For that reason, we analyzed further the results obtained in these two crop species.

The results from moss and melon differ in the number of TE families responsible for the PM-TE observed. In melon, nine families that occupy almost 4% of the genome are responsible of 36% PM-TE. These diverse families comprise relative young element, suggesting a TE recent activity in melon genome. In the case of *P. patens*, the vast majority of PM-TE are due to a specific *Gypsy* family, RLG1, consistent with the fact that this family represents a quarter of the moss genome.

The distribution of PM-TEs with respect to genes differs in melon and in moss genomes. About 22% of PM-TE in melon and about 3% in moss are located within genes. This can be explained by the TE distribution across the chromosomes in these two genomes. Whereas TEs in the moss genome are dispersed along chromosomes interleaved with genic regions (Rensing et al. submitted), melon chromosomes present two distinct regions (see Chapter 1). Indeed, most of the PM-TE in moss are due to a specific *Gypsy* family, RLG1, which encodes an integrase that contains a chromodomain. Chromodomain usually mediates interactions with heterochromatin, suggesting that RLG1 elements target heterochromatin for insertion. Although few PM-TE are located within genes, these polymorphic TE insertions can still have an impact on the evolution of the moss genome. On the other hand, the melon analysis presents an important fraction of PM-TE located in genic regions and maybe associated with interesting traits.

In order to correlate PM-TE to particular agronomic traits, we focused on PM-TE that are polymorphic between the two elite melon varieties PS and VED. These two melon elite lines are closely related phylogenetically, but they differ in many important agronomical traits, such as fruit shape, flesh color, sugar content and aromas (Sanseverino et al. 2015).

The GO enrichment analysis and getting more information about the gene function allowed us to end up with a list of genes related to flower development. The results obtained suggest that the TE insertions, located within its candidate gene and polymorphic between the two elite lines, did not altered the expression of the gene in the tested conditions. More analysis should be performed to determine the impact of TE insertion in these candidate genes. However, this information is a great source for studying the genetic variability due to TE insertion polymorphisms and for detecting genomic regions involved in domestication.

CHAPTER 3

Transposable elements impact on gene regulation

CHAPTER 3: Transposable elements impact on gene regulation

The E2F transcription factor binding sites amplified by MITEs during evolution of *Brassica* species

3.1.- INTRODUCTION

Apart from being involved in chromosome structure, TEs can be inserted within or close to genes and can impact them in many different ways. The most common effect of inserted TEs within or close to a gene is its inactivation. But TEs may have other effects on gene regulation that should be considered. The fact that TEs are targets of silencing makes them able to modify gene expression by attracting new combinations of epigenetic marks to nearby genes (Ahmed et al. 2011). TEs contain their own promoters, terminators and regulatory signals, which can also provide novel alternative promoters, terminators or splice sites leading to new expression patterns (Cowley and Oakey 2013; Lisch 2013).

Thanks to the bioinformatics era, it has been possible to identify conserved regulatory sequences in eukaryote genomes, showing that some transcription factor binding sites (TFBSs) may co-localize with TEs. In humans, there are several examples where TEs are associated with the binding sites of TFBS, such as p53, POU5F1, SOX2, c–Myc, CTCF, OCT4, NANOG and ERa (Wang et al. 2007, 2009; Bourque et al. 2008; Bourque 2009; Kunarso et al. 2010; Lynch et al. 2011; Schmidt et al. 2012; Jacques et al. 2013). This fact can raise the idea that TEs may have the capacity to generate new transcriptional networks. Their characteristic mobility makes them a key on relocating binding sites across the genome (Gifford et al. 2013).

This project started with the interest to annotate in a more precise way miniature inverted-repeat transposable elements (MITEs) in *Arabidopsis thaliana*. While analyzing the results, a peculiar MITE came across which contained short repeated sequences. This sequence was fitting the consensus for the E2F binding site (E2F BS) (TTssCGssAA, where s = C or G; Ramirez-Parra et al. 2003; Vandepoele et al. 2005).

E2F proteins are a family of transcription factors that regulate the expression of genes involved in cell cycle, DNA replication, DNA repair, cell proliferation, differentiation and development (Ramirez-Parra et al. 2007; van den Heuvel and Dyson 2008; Lammens et al. 2009; Biswas and Johnson 2012). E2F BS is a well-conserved motif involved in crucial functions as DNA replication and cell cycle (De Veylder et al. 2002; Ramirez-Parra et al. 2004). E2F TFBSs are evolutionary well conserved either in animals or plants and all members of this TF family recognize the same consensus sequence in order to bind DNA (DeGregori and Johnson 2006). In mammals, a total of eight E2F proteins were characterized, while in *A. thaliana* six (Lammens et al. 2009). In *A. thaliana,* E2F BS have been found in promoter regions of genes involved in DNA repair and chromatin dynamics, such as CDC6, MCM3, ORG1, CDTa, PCNA, RBR and RNR (Naouar et al. 2009; Vandepoele et al. 2005).

Dr. Elizabeth Hénaff, past member of the lab, looked for E2F BS sequences in the available annotation of TEs in *A. thaliana* (Ahmed et al. 2011), she found that 73% of E2F BS were within an annotated TE (we called these TEs as E2F-TE). This result is much higher than what it would be expected for a random distribution, because the TE fraction accounts for 21% of the *A. thaliana* genome (Ahmed et al. 2011). Moreover, analyzing the distributions of other well-known plant TFBS, none of them are found in TEs at a proportion higher than expected for a random distribution. These analyses suggested that TEs have amplified the E2F BS in Arabidopsis.

E2F BS sequences were studied in the TE and non-TE fraction of the genome. Among all sequences fitting the E2F BS consensus only the sequence TTCCCGCCAA is found at much higher number compared to the other sequences (at least 14 times more). Furthermore, 90% of this sequence is located in TEs, suggesting that TEs have amplified the sequence TTCCCGCCAA in *A. thaliana*.

3.2.- OBJECTIVES

Previous bioinformatic analyses suggested that the E2F binding site has been captured and amplified by MITEs in several *Brassica* species. The goal of this project is to assess the impact of E2F-TEs in reprogramming gene regulation on the E2F transcriptional network. This objective is divided in the following ones:

- Study the chromatin structure in which E2F-TEs are located in TEs and outside TEs
- Determine whether E2F-TEs have the capacity to bind E2F transcription factor *in vivo*

3.3.- MATERIALS AND METHODS

Plant material

Plants were grown on soil at 22ºC under long-day (16 hours light / 8 hours dark) photoperiod. The Columbia (Col) ecotype of *Arabidopsis thaliana* was used as wild-type. E2Fa-DPa OE seeds were obtained from Dr. Crisanto Gutierrez (CBMSO, Madrid, Spain). The *Arabidopsis thaliana* accessions Ler-1, Bur-0, C24 and Kro-0 were obtained from Arabidopsis Biological Resource Center (ABRC), having stock numbers CS22686, CS22679, CS22680 and CS1301, respectively.

Chromatin immunoprecipitations

Chromatin immunoprecipitation analysis (ChIP) was performed as previously described in Bowler et al. (2004), with some modifications. The aerial part of 17-days old E2Fa-DPa^{OE} and wild-type was used as starting material. Crosslinking was performed by vacuum infiltration in 37% formaldehyde buffer for 15 minutes. Chromatin was isolated and sheared by sonication with 10 seconds pulses for 3 times (10’’ on 5’’ off at 10% amplitude). Chromatin was immunoprecipitated using 100 ul of sonicated chromatin and the antibodies anti-monomethyl histone H3 (Lys27) (H3K27me1, Upstate Millipore, http://www.millipore.com/, reference 07-448), anti-dimethyl histone H3 (Lys4) (H3K4me2, Upstate Millipore, http://www.millipore.com/, reference 07-030). The E2Fa antibody was given by Dr. Lieven De Veylder (VIB Department of Plant Systems Biology, University of Ghent, Belgium). For E2Fa immunoprecipitations, Low-cell ChIP kit (Diagenode, http://www.diagenode.com/) was performed obtaining indistinguishable results. As negative controls, rabbit IgG (Diagenode) or no antibody were used for immunoprecipitations performed using the Low-cell ChIP kit (Diagenode) or the standard method, respectively. At least two biological replicates were performed for all ChIP experiments.

PCR analyses

Immunoprecipitated DNA was analyzed by semi-quantitative PCR. And for TE insertion polymorphisms analysis, young leaves samples were used for DNA extraction (Kasajima et al. 2004).

Primer pairs were designed to be 20-26 bp long for PCR amplification using Primer 3 software (Untergasser et al. 2012). Primers were coded as ‘Ath_Hat1/Hat2/Guy1_X’, where the acronym stays for *Arabidopsis thaliana* followed by the TE superfamily (Hat1 as SimpleHat1, Hat2 as SimpleHat2 and Guy1 as SimpleGuy1), and X being the TE’s copy number. The oligonucleotides used in PCR amplifications are listed in Table 3.1.

PCR products were obtained in a total volume of 20ul using 20uM of each primer, 0.25uM of dNTPs and LongAmp® Taq DNA Polymerase (New England BioLabs), according to manufacturer’s instructions. Half of the PCR product volume was separated on a 1% agarose gel and stained with ethidium bromide for checking the PCR amplification. Fragment sizes were estimated with the 1kb DNA ladder (Biotools).

Table 3.1. List of oligos

Full name	Name	Sequence	Product size (bp)	Number of E2F	Distance to closest gene (bp)
Ath_TE_12952_SIMPLEHAT2	AtHat2_12952 AtHat2_12952-r	GAAGAGAGTGAAGAACGGAGGA CGTTGAAAGTCGGTAAAAATCC	1075	---	---
Genomic-region_chr4_3724795-3727015	At_genomic-region1 At_genomic-region1-r	CTTGACATACTTGAGGAACCGAC AAGATTTAGAGATGGAGAATTGGCC	743	---	---
EXPANSIN 3	EXP3-f EXP3-r	TTGCCACCTTCGGTTTAGTC TAGAAAGTGGCGTGTGCATT	354	---	---
EXPANSIN 7	EXP7-f EXP7-r	CCCTGACATTCTCTCCCAAA ATAAGTTGACGTGCGAGCAG	350	---	---
MINICHROMOSOME MAINTENANCE 5	MCM5-f MCM5-r	CTGACATCGTTGCTTCGTCTC GGAATTGAAAATGCTTACAACG	382	---	---
PROLIFERATING CELLULAR NUCLEAR ANTIGEN 1	PCNA1-f PCNA1-r	CTAGGGCAAAGTCGGTTTTGG AGCTCCAACATTTCGTCGTC	352	---	---
AtSimpleHat1_E2F_18_borders_239_304_copy_0	AtHat1_0-f AtHat1_0_in1-r	CCCAGTGGGCATTAAAGAGA TCGGGAAAAAGGTTGAATTGC	722	6	47
AtSimpleHat2_borders_copy_74	AtHat2_74-f AtHat2_74_in1r	TCGGGAGGATGATGTTTAGG TTTTTGCGGGAAGATTATGG	690	5	318
AtSimpleGuy1_tagE2F_92|84r_borders_B_copy_6	AtGuyB_6-f AtGuyB_6_in1-r	GAGTCAGACTTGTCTCGCGTAA CATATTTTGCTGTTTTGGCAAG	768	5	0
AtSimpleGuy1_tagE2F_83_borders_A_copy_7	AtGuyA_7-f AtGuyA_7_in1-r	CGAAGGGAACATTCACTTTACA CGGAAGAACATAATTTTTGTGG	815	5	115
SimpleHat2_borders_copy_52	AtHat2_52-f AtHat2_52-r	AAATATACAAGCGATGAAATTGAGAA CAGAAGATTTTGTTTTACCCAAGC	991	11	558
SimpleGuy1_tagE2F_92l84r_borders_B_copy_25	AtGuy1_25-f AtGuy1_25_r2-r	AGAGGAATTAGACCAAAGAGCAGA CCGTCAAGAACAGAATCTCGTAG	1882	10	0

Full name	Name	Sequence	Product size (bp)	Number of E2F	Distance to closest gene (bp)
SimpleGuy1_tagE2F_83_borders_A_copy_9	AtGuy1A_9_f2-f	AAACACCCAATTACATCAGCAAC	1614	6	834
	AtGuy1_9-r	CGAAACCCACGTTTAGTGAATCA			
SimpleHat2_borders_copy_4	AtHat2_4-f	AACTTGTAGAAAGGCGACAGTTG	2070	9	721
	AtHat2_4-r	GTGAAGCCGTGAGATTTCTTCT			
SimpleHat2_borders_copy_28	AtHat2_28-f	CGAATAAGATTCAACTGTTCATGC	1815	10	319
	AtHat2_28-r	AGGTAGAGTTATGGGAACTTGTCG			
SimpleGuy1_tagE2F_83_borders_A_copy_6	AtGuy1A_6-f	ACGTGATCTGAAATGTTGGTCTAA	1627	5	0
	AtGuy1A_6-r	CTCCAGAGTCTTTGATCTACCGTT			
SimpleHat2_borders_copy_87	AtHat2_87-f	TGTACAAGCGATGAAATTGAGAA	849	5	972
	AtHat2_87-r	GAATCGTACGTCTCTTTTTGGAA			
SimpleHat2_borders_copy_78	AtHat2_78-f	TTTCGGGTTTAAGCTTTTCG	739	7	930
	AtHat2_78-r	GTCATACCATACCATCCATGCTT			
SimpleGuy1_tagE2F_92l84r_borders_B_copy_20	AtGuy1_20-f	GGAACAATACTCAGCCCTGTTT	829	6	827
	AtGuy1_20-r	TGGTCCATCTGAATGACTTTGT			
SimpleGuy1_tagE2F_92l84r_borders_B_copy_18	AtGuy1_18-f	AATTGTATTCATTTTCCCGTCAA	812	10	536
	AtGuy1_18-r	AGACCTGACACCAAGACCAAGTA			
AtSimpleHat1_E2F_18_borders_239_304_copy_2	AtHat1_2-f	TGTATCGTGTGTAAAGATCTTGGT	848	5	2305
	AtHat1_2-r	GTACCCAACTTGGTGTTTGTCAT			
AtSimpleHat1_E2F_18_borders_239_304_copy_3	AtHat1_3-f	AAATCCTTTTCTTGGTCGGAAAT	807	7	1683
	AtHat1_3-r	TGATTCGTTAGATTCGTTGAACA			
SimpleHat2_borders_copy_49	AtHat2_49-f	CATCACCTATGGAGAAGTTGGAG	1148	15	2519
	AtHat2_49-r	TTTACGGATTCCACTTTTTATGG			
SimpleGuy1_tagE2F_92l84r_borders_B_copy_22	AtGuy1_22-f	ATTTTACGTCATTCGTTTTTCCC	998	11	3268
	AtGuy1_22-r	GTCCAAGGATGAGCATTGAGAGTTG			
SimpleGuy1_tagE2F_92l84r_borders_B_copy_12	AtGuy1_12-f	TACGGATTTGTGAAAACATGATG	903	10	5978
	AtGuy1_12-r	TATCAACAATGGGGTTCATCCTC			
SimpleHat2_borders_copy_64	AtHat2_64-f	GGTTTTAGATAGTTTACCCGCACTA	943	9	31516
	AtHat2_64-r	GTTTACCCAATTAACCCATCAAG			
SimpleHat2_borders_copy_10	AtHat2_10-f	ACCCAATTAACCCATCAAGTTTTG	995	3	5390
	AtHat2_10-r	AGCACTGTCTCGGTTCTTCATAAG			
SimpleHat2_borders_copy_13	AtHat2_13-f	ACTAAGTTTGGGGTGGAATTGGC	1103	10	12741
	AtHat2_13-r	ATCACAGGTTTACAGGTTTACGT			
SimpleHat2_borders_copy_31	AtHat2_31-f	CAGATTCCACTTTTTACGGGTTTACG	964	7	5064
	AtHat2_31-r	AAGCCACTATTGCTGGTACTGTG			
SimpleGuy1_tagE2F_92l84r_borders_B_copy_21	AtGuy1_21-f	GCTTGTAATAGCAACGATGACAC	968	6	7258
	AtGuy1_21-r	TCATAATTTTCCTAGTAAACCGCA			
SimpleHat2_borders_copy_66	AtHat2_66-f	GTCAAAATGGGTAATCCAACTCA	924	1	5083
	AtHat2_66-r	CAATTTGATAATCCAACCTAGCAA			
SimpleHat2_borders_copy_88	AtHat2_88-f	TGAAGCCTAAGTTTACCCTCACA	860	1	16498
	AtHat2_88-r	CCTCAGACTCTCCTCGACACAAA			

3.4.- RESULTS

Previous results showed that TEs have amplified E2F BSs in various *Brassica* genomes. Most of the E2F-TEs are located far from genes, but there is a fraction of them which are close or within genes. It is known that the chromatin context may influence the accessibility of TFs to their binding sites. Therefore, we decided to investigate the chromatin in which E2F-TEs are located. A preliminary analysis performed by Dr. Elizabeth Hénaff using a whole-genome ChIP-seq data for the histone modifications H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K9Ac and H3K9me2 (Luo et al. 2013) showed that H3K4me2 is associated with E2F sites outside TEs, while H3K27me1 is linked with E2F sites inside TEs.

In order to capture those states and to study the epigenetic marks actually associated with E2F in TEs, we designed a Chromatin Immuno-Precipitation (ChIP) experiment analyzing the samples by semi-quantitative PCR. Due to repetitiveness nature of TEs, one primer needs to be designed in the non-repetitive region flanking the E2F-TE element in order to ensure specific amplification, and the second primer could be designed within the element itself and including E2F TFBSs. We arbitrarily considered E2F-TEs found at more than 1kb as far from genes, and those laying at less than 1kb from an annotated gene as close. A total of 26 E2F-TEs were analyzed (Table 3.2) and four known E2F target genes in at least two independent ChIP experiments. As a negative control, we analyzed a TE which belongs to *SimpleHat2* family, which is a MITE family that contains E2F TFBS, but this element has diverged enough that not contains E2F TFBS motifs. We used a genomic sequence located 4kb upstream of an E2F-TE as an additional negative control.

Table 3.2. Summary of genomic location and epigenetic context of the analyzed E2F-TEs.

Distance to closest gene	# E2F-TEs	# E2F BS	Analyzed by PCR	H3K4me2-rich	H3K27me1-rich
less than 1Kb	109	663	14	5	8
greater than 1Kb	99	570	12	0	12
Total	208	1233	26	5	20

The results confirmed that E2F BSs positioned in the promoter region of E2F regulated genes are associated with a high level of H3K4me2 and a low level of H3K27me1 (Figure 3.1).

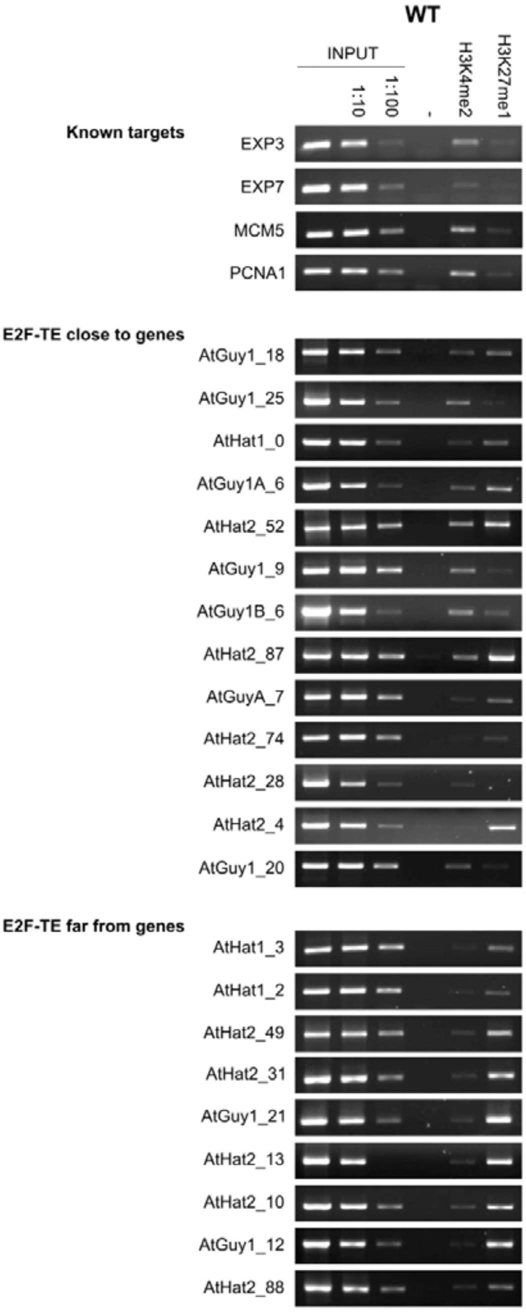


Figure 3.1. ChIP analyses of the epigenetic marks (H3K4me2 and H3K27me1) associated with different types of E2F BS. Negative control (-) was performed immunoprecipitating with no antibody.

On the contrary, E2F-TEs located far from genes showed high level of H3K27me1 and a low level of H3K4me2 (Figure 3.1). E2F-TEs inserted close to or within genes showed higher levels of H3K4me2 and lower levels of H3K27me1 than those located far from genes, and some of them showed a high H3K4me2 and a low level of H3K27me1 similar to the control E2F sitting in gene promoters (Figure 3.1). This suggests that the E2F-TEs located close to genes tend to be associated with open chromatin marks whereas those far from genes are associated with heterochromatin.

Having checked the chromatin state of those E2F-TE elements, we analyzed whether the E2F TFBS protein can actually recognize and bind the E2F-TE sites. ChIP analyses were performed in order to know whether E2F binds *in vivo* to various E2F BSs located in TEs across genome. A specific antibody against E2Fa TF (Heyman et al. 2011) was used. We analyzed the same 26 E2F-TEs, the four positive and negative controls in at least two independent ChIP assays. Our results (Figure 3.2) show that E2F binds all the positive control E2F BS, whereas no binding is detected for the negative control sequences (Figure 3.2).

We detected binding of E2Fa *in vivo* to all studied E2F-TEs irrespective of them being close or far from genes (Figure 3.2). In all cases, the intensity of the amplified band was increased in plants over-expressing the E2Fa-DPa factors (De Veyler et al. 2002), confirming the specificity of the binding signal detected and suggesting that the E2F protein is limiting for the binding.

Our results suggest that the E2F-TEs contribute to the E2F BS in *A. thaliana*. Therefore, we have analyzed if these E2F-TEs, or at least those located close to genes, are fixed in Arabidopsis.

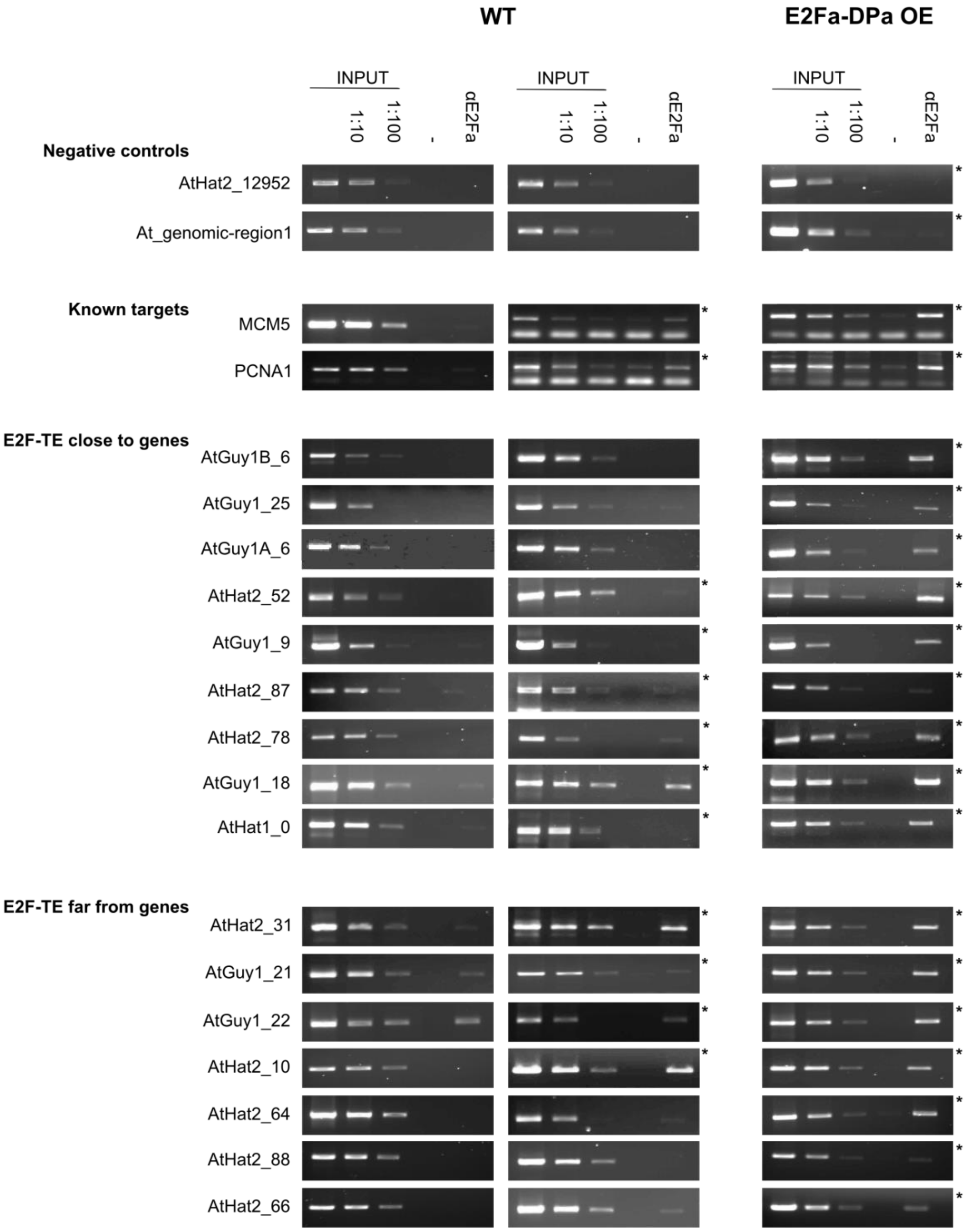


Figure 3. 2. ChIP analyses of the binding of the E2F protein to E2F BS. Two different ChIP analyses in wild type plants are shown, together with the analysis in plants over-expressing the E2Fa-DPa proteins. Negative control (-) was performed immunoprecipitating with no antibody or with an anti-IgG antibody (where indicated by *).

We took advantage of the available genome assemblies of 4 *A. thaliana* ecotypes (Ler-1, Bur-0, C24 and Kro-0) (Schneeberger et al. 2010) to check whether E2F-TEs close to genes are fixed among *A. thaliana* ecotypes. Dr. Elizabeth Hénaff used as queries seven E2F-TEs plus flanking gene sequences to perform a BLAST against each assembled genome in order to find the orthologous locus. PCR analyses were performed to confirm the presence or absence of the E2F-TE in each ecotype at a given locus (Table 3.3).

Table 3. 3. E2F-TE insertion polymorphisms among different ecotypes of *A. thaliana*. The accession number of the gene closest to each E2F-TE insertion is given, indicating the distance between gene and E2F-TE and where E2F-TE is located respect to the gene. PCR amplification of each locus revealed the presence (+) or the absence (-) of the corresponding E2F-TE. Failure to amplify the corresponding locus in a particular ecotype is shown by N.D. (not determined).

E2F-TE insertion	**Closest gene to E2F-TE insertion**	**Distance to gene (bp)**	**Location respect to gene**	**Col-0**	**Bur-0**	**C24**	**Kro-0**	**Ler-1**
AtHat2_28	AT1G66780	609	5 prime	+	-	-	-	-
AtHat2_4	AT1G22590	796	3 prime	+	-	-	-	+
AtGuy1A_6	AT3G53310	330	3 prime	+	-	+	+	+
AtGuy1A_9	AT5G24670	834	5 prime	+	+	N.D.	+	-
AtGuy1_25	AT3G61020	0	intra	+	N.D.	+	N.D.	-
AtHat2_38	AT2G15400	278	3 prime	+	+	N.D.	N.D.	+
AtHat2_52	AT3G29810	558	5 prime	+	N.D.	N.D.	N.D.	N.D.

We were able to detect polymorphic E2F-TEs in five out of seven analyzed cases (Figure 3.3). For several locus and ecotypes, we couldn’t neither get the sequence nor design primers to verify the presence or absence of the E2F-TE.

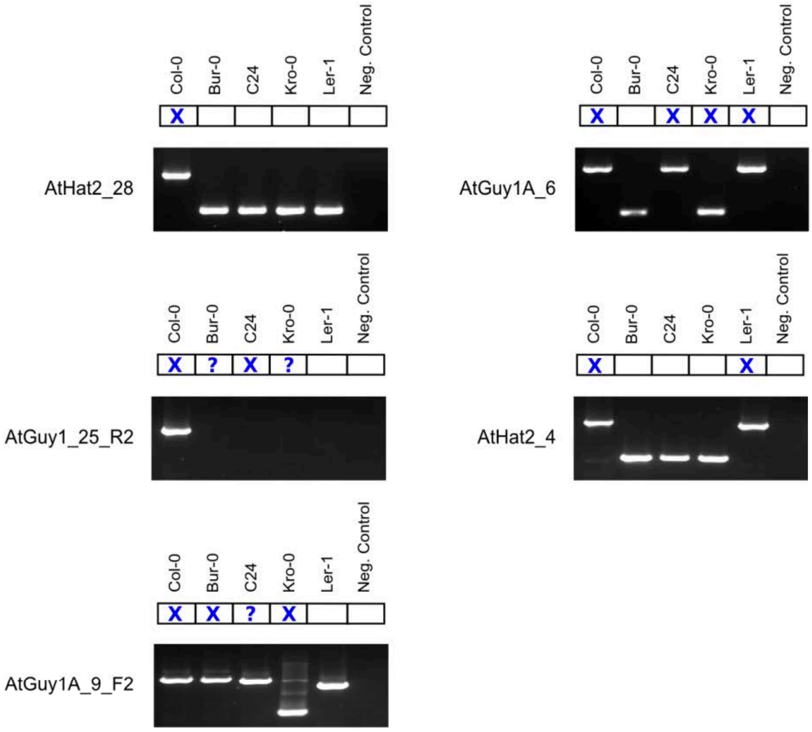


Figure 3.3. PCR validations of TE insertions polymorphic among the *A. thaliana* ecotypes analyzed. The name of the ecotype is indicated at the top. A blue cross indicates ecotypes estimated to contain the TE insertion, and a question mark indicates that BLAST analysis couldn't determine the presence or absence of the E2F-TE in a given locus.

For the ones confirmed by PCR, we were able to obtain the sequence and to confirm the presence or absence of E2F-TE in the *A. thaliana* ecotypes (Figure 3.4).

This result suggests that TEs containing E2F TFBSs have transposed recently, resulting in different genes potentially wired into the E2F transcriptional network in the different analyzed ecotypes.

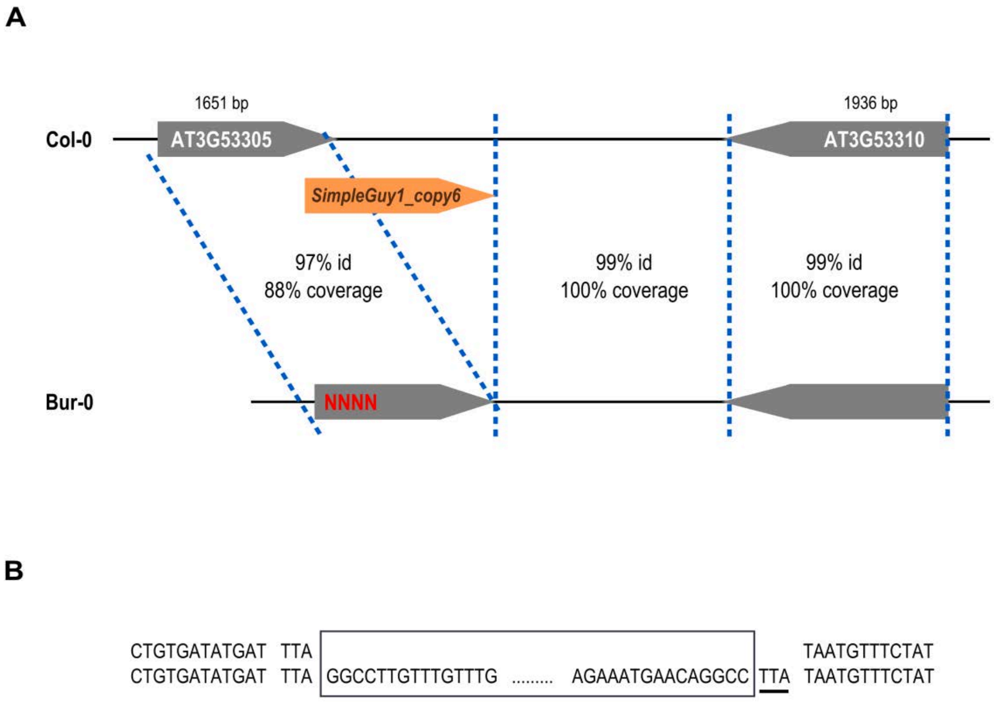


Figure 3. 4. Example of an E2F-TE polymorphic between Col-0 and Bur-0 *A. thaliana* ecotypes. The presence of a *SimpleGuy1* family element in Col-0 and the absence in Bur-0 is represented. The percentage of coverage and identity is indicated in the different compartments of the loci (A). The 5 prime and 3 prime flanking sequences are shown with part of target inverted repeat (TIR) of the E2F-TE element present in Col-0 (boxed). The target site duplications are underlined (B).

3.5.- DISCUSSION

Past members of the group demonstrated that the E2F BS has been captured and amplified by several MITE families. The capture of E2F BS happened in an ancestral *Brassica* genome by these MITE families.

The results presented here show that E2F BSs within TEs are bound by the E2F protein *in vivo*. The ChIP analyses showed that the E2F TF binds to E2FBS within TEs (E2F-TEs) irrespective to their location with respect to genes. Although all cases presented some variability from experiment to experiment in the binding level, when the experiment is performed using plants over-expressing the E2F transcription factor E2Fa-DPa, there's a clear increase in the binding for all those analyzed sites. These observations suggest that all E2F BS can be bound by the E2F TF and that the concentration of the E2F factor is limiting for the binding and, therefore, the amplification of E2F BSs may have affected the binding of E2F to its sites.

Moreover, it's well known that TEs are associated to heterochromatic marks, being H3K27me1 associated with heterochromatin and silencing in *Arabidopsis* (Jacob et al. 2009). ChIP analyses of histone marks allow us to examine the association between E2F-TEs and certain epigenetic marks, knowing their position respect to genes. We could observe that isolated E2Fs are associated with euchromatic marks, while most of analyzed E2F-TE cases are linked to heterochromatic marks (high levels of H3K27me1 and low levels of H3K4me2), except for those E2F-TEs that are within or close to genes (high levels of H3K4me2 and low levels of H3K27me1).

As shown before, some E2F-TEs are found within or close to genes. These genes are putative E2F targets, and interestingly the closest E2F BS is within a TE. We have therefore analyzed different ecotypes of *A. thaliana*, in order to know whether E2F-TEs are fixed in the population. Surprisingly, we find out some polymorphic E2F-TEs among those ecotypes, suggesting that those elements have been transposed recently. This suggests that the E2F transcriptional regulatory network may differ in different ecotypes.

However, it is not straightforward to anticipate the effect of an insertion of an E2F-TE close to a gene as the E2F factors can function both as activators or repressors of transcription, depending on the cellular context and target gene (Biswas and Johnson 2012). Further studies should be performed to know the biologic effects of including certain genes into the E2F transcriptional network.

Depending on the location of E2F-TEs respect to genes, they may impact at different levels in the regulation of E2F transcriptional network. On one hand, E2F-TEs located close to genes may directly participate in gene promoters and it’s known that some genes are regulated by E2F, like three expansin genes (Ramirez-Parra et al. 2004). These E2F-TEs close to genes may have a direct effect incorporating new genes into the E2F transcriptional network, as examples the polymorphic E2F-TE among *A. thaliana* accessions.

On the other hand, the vast majority of E2F-TEs are found far from genes, suggesting that these elements don’t regulate directly genes, but still they have an effect on gene regulation. Maybe E2F-TEs found far from genes are able to bind E2F and they may reduce the E2F proteins available to bind the E2F-TEs in gene promoters. Another possibility is that E2F-TEs located far from genes maintain the capacity to be mobilized during evolution and they can be a reservoir to rewire new genes into the E2F transcriptional network.

CHAPTER 4

Use of retrotransposon-based vectors to introduce variability into the genome

CHAPTER 4: Use of retrotransposon-based vectors to introduce variability into the genome

High efficiency transposition of the tobacco retrotransposon Tnt1 in *Physcomitrella patens*

4.1.- INTRODUCTION

The moss *Physcomitrella patens* is a non-vascular multicellular land plant and a member of the bryophyte family which diverged from the other land plant lineages more than 400 Mya (Kenrick and Crane 1997). This species has become a model system to study plant development, growth and cell differentiation (Sakakibara et al. 2003; Repp et al. 2004). One of the interesting features of *P. patens* is the high frequency of homologous recombination (Schaefer and Zryd 1997), which facilitates molecular genetic approaches to study gene function in plants. Another interesting characteristic of *P. patens* is the dominant haploid phase of its life cycle. These characteristics make *P. patens* a suitable system for reverse genetic approaches allowing to study the relation between moss organogenesis and many regulators of plant development including growth factors, the cytoskeleton, transduction pathways, transcription factors, epigenetic control and dedifferentiation processes (Bonhomme et al. 2013).

However, *P. patens* also has some disadvantages as a study model. The high efficiency of homologous recombination is accompanied by a very low integration efficiency for sequences without any similarity to the genome of moss *P. patens* (Schaefer and Zryd 1997) which makes it difficult to develop mutant collections based on the insertion of DNA (e.g. *Agrobacterium* mediated T-DNA insertions).

In vascular plants, forward and reverse genetic approaches have been implemented using DNA transposons or LTR retrotransposons from endogenous or heterologous species (Sundaresan 1996; Wisman et al. 1998; Meissner et al. 2000). Due to their nature, LTR retrotransposons create stable mutations once inserted and, interestingly, many LTR retrotransposons tend to insert within gene-rich regions (Okamoto and Hirochika 2000; Le et al. 2007; Urbanski et al. 2012), making them suitable for gene tagging. Several examples where LTR retrotransposons were used to generate mutant collections have been published. For instance, the rice Tos17 (Hirochika 2001) or the *Lotus japonicus* LORE1 elements (Fukai et al. 2012; Urbanski et al 2012) were used to generate a mutant collection in their respective species, whereas the tobacco (*Nicotiana tabacum*) Tnt1 element was employed in heterologous hosts such as Arabidopsis, *M. truncatula,* lettuce and soybean (Lucas et al. 1995; D'Erfurth et al. 2003; Mazier et al. 2007; Cui et al. 2013). Tnt1 is an autonomous 5.3 kb long *Copia*-like LTR element which generates a 5 bp duplication after its insertion in the genome. Among plant retrotransposons, Tnt1 is one of the best characterized and its transcription and transposition can be easily induced (Grandbastien et al. 2005).

This part of this PhD work was done in collaboration with the group of Dr. Fabien Nogué and Dr. Marie-Angèle Grandbastien (INRA AgroParisTech – IJPB, France). Previous results had shown that Tnt1 can also transpose efficiently in *P. patens*. Isolated protoplasts from *P. patens* were transformed with the plasmid Tnk23 containing the entire Tnt1-94 copy (Grandbastien et al. 1989) together with a kanamycin resistance gene (Lucas et al. 1995). PCR analysis of 18 independent clones showed that Tnt1 had integrated into the genome, but no other plasmid sequence was found to be integrated. This result suggested that Tnt1 inserted into the genome by retrotransposition from a non-integrated copy of Tnt1.

Sequence-Specific Amplification Polymorphism (SSAP) (Waugh et al. 1997) analysis was performed to determine the number of Tnt1 integrated sequences. The results suggested that the number of integrated Tnt1 in each clone varied from 1 to 10 copies. 22 SSAP bands were cloned and sequenced, demonstrating that Tnt1 inserted into the genome of *P. patens* by retrotransposition mechanism. In 54% of the cases, Tnt1 inserted within a gene and 77% at less than 1 kb from a gene. These results are similar to what has previously shown in its natural host (Le et al. 2007). However, in order to be a useful genetic tool, the insertions of Tnt1 should be stable, allowing the characterization of stable phenotypes. To this end, the inserted Tnt1 elements should not be able to move anymore.

Therefore, the objectives of this work were to study whether inserted Tnt1 maintains the capacity to transcribe and transpose again. The final goal is to obtain a Tnt1-based system allowing to produce a collection of stable mutants and to get a useful tool for forward genetics in *P. patens*.

4.2.- OBJECTIVES

The main objective of this chapter is the development of a genetic tool based on the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon for an efficient insertion mutagenesis in *Physcomitrella patens*. This goal can be divided in the following points:

- Analyze whether the Tnt1 insertions in *P. patens* are stable.
- If not, develop a useful Tnt1-based vector system to create stable mutant collection.

4.3.- MATERIALS AND METHODS

Plant material

P. patens (Hedw.) B.S.G. 'Gransden2004' was vegetatively propagated as previously described (Cove et al. 2009). For the isolation of protoplasts, protonemal filaments were cultured from subculture of homogenized tissue on BCD agar medium supplemented with 1mM $CaCl_2$ and 5 mM ammonium tartrate (BCDAT medium) overlain with cellophane.

Bacterial strains and constructs

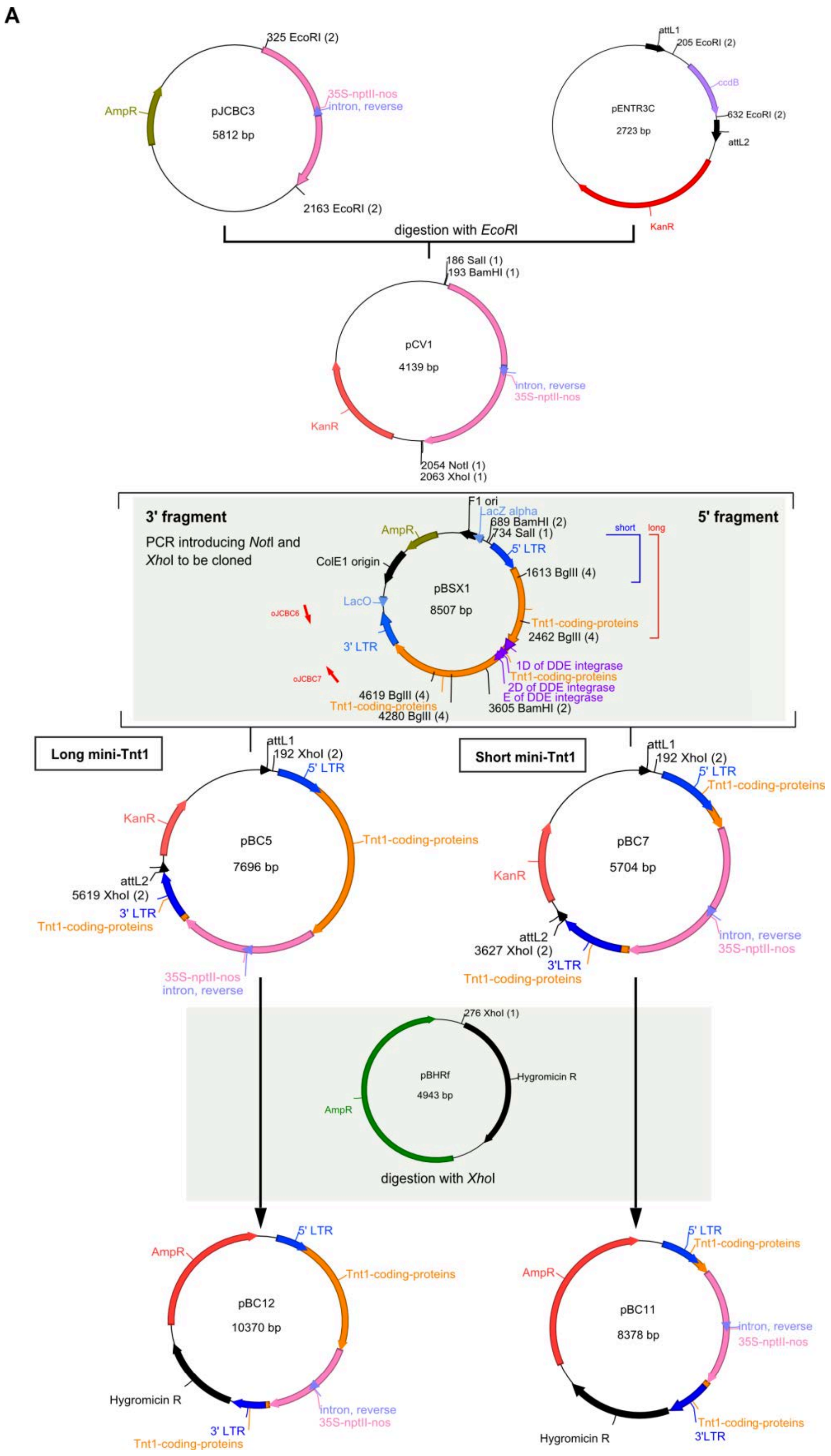
The tagged mini-Tnt1 element is a derivative of the pBIN19 vector containing the Tnt1-94 retrotransposon element from tobacco (X13777) and has been constructed as follows. A double stranded oligonucleotide corresponding to a previously described artificial intron (Hou et al. 2010) was cloned into the *Msc*I site of the pBNRf plasmid containing an *nptII* expression cassette (Schaefer et al. 2010) and a clone, pJCMN21, with the intron in reverse orientation with respect to the *nptII* cassette was selected. The interrupted cassette was amplified by PCR with the oJCBC4 and oJCBC5 primers (Table 4.1) and was cloned into the pCRII plasmid (Invitrogen), obtaining pJCBC3. An *EcoR*I fragment of pJCBC3, containing the *nptII* interrupted cassette, was cloned into pENTR3C (Invitrogen). The 5' fragments of Tnt1 were obtained from pBSX1 (Lucas et al. 1995), by digestion *Sal*I-*BamH*I (long mini-Tnt1) or *Sal*I-*Bgl*II fragment (short mini-Tnt1), whereas the 3' fragment was amplified by PCR with the oJCBC6 and oJCBC7 primers (Table 4.1) on the pBSX1 plasmid. Both fragments were cloned into the pENTR3C plasmid containing the interrupted *nptII* cassette to give the pBC5 (long mini-Tnt1) and pBC7 (short mini-Tnt1) plasmids. The *XhoI* fragments of pBC5 and pBC7, which contain the complete mini-Tnt1 elements, were cloned into the *XhoI* site of the pBHRf vector containing a hygromycin resistance cassette (Schaefer et al. 2010) to obtain the pBC12 (long mini-Tnt1, with an element of 5,325 nt) and the pBC11 (short mini-Tnt1, with an element of 3,420 nt) plasmids.

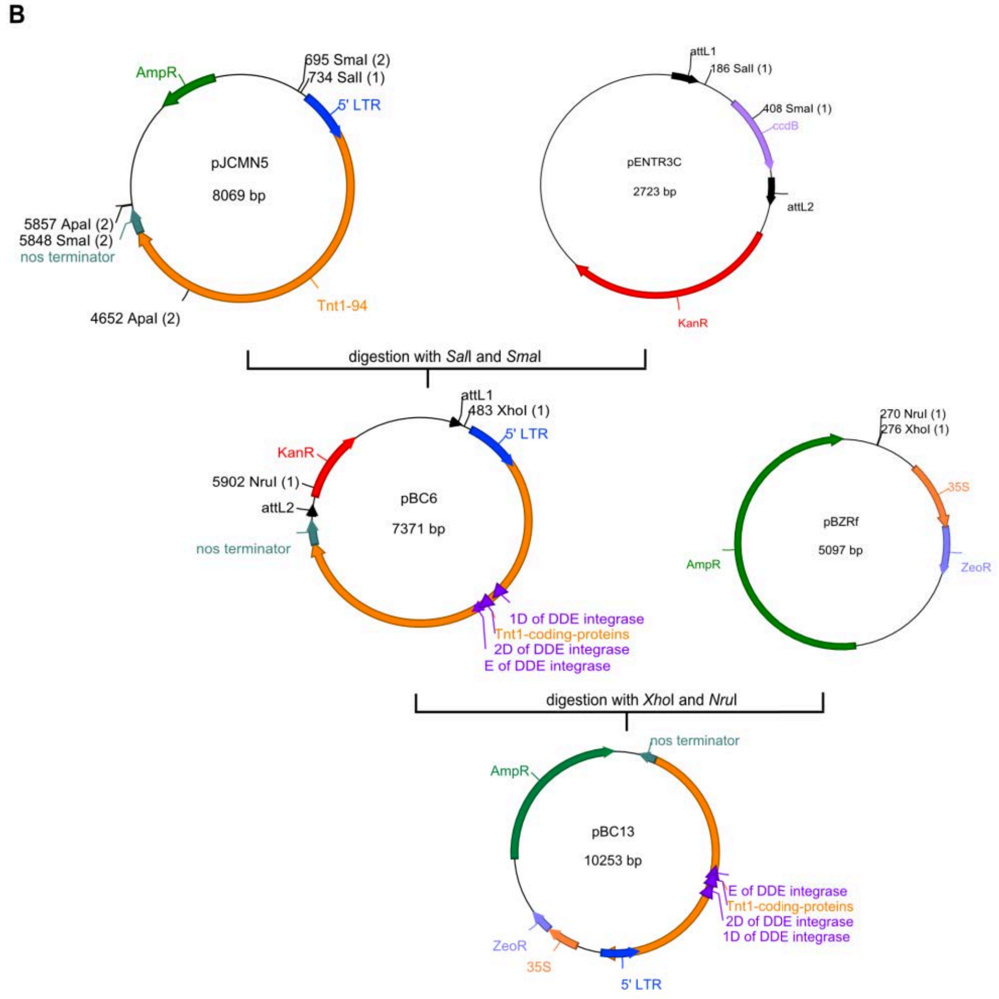
For the transient expression of Tnt1 proteins, two types of vectors were created. The *Apa*I fragment of a plasmid containing nos terminator was cloned into *Apa*I of pBSX1 (Lucas et al. 1995), eliminating the Tnt1 3' LTR, and obtaining the pJCMN5 plasmid. The vector pBC6 expressing the wild type Tnt1 proteins under the control of the Tnt1 5' LTR and nos terminator was obtained by cloning the *Sal*I-*Sma*I fragment of pJCMN5 into *Sal*I-*Sma*I sites of the pENTR3C plasmid. Another protein construct was created mutating the second amino acid D from the integrase DDE domain. The mutated integrase, changing D to A, was amplified by PCR using a combination of four primers: oJCMN13, oJCMN14 (which includes the mutation), oJCMN15 (which includes the mutation) and oJCMN16 (Table 4.1).

Table 4.1. List of oligos used in this study

Usage of primers	Oligo name	Sequence
Mini-Tnt1 integration detection	oJCMN15	GGCACAAAAGAATGGGTCATATG
	oJCMN16	CCAACTGCTCCACTTCAAGATC
	oJCMN23	GGTGGAGAGGCTATTCGGCTATG
	oJCMN24	GCAGGAGCAAGGTGAGATGACAG
	oBC1	CAGGTTCTGCTCGTTCACTG
	oBC2	ATCTCCCCCTCCAGTCTCAT
	oCV1	GCTTTCAGCTTCGATGTAGGAG
	oCV2	AGAAGAAGATGTTGGCGACCT
	APTg-f	TAGGGTTGCTTTCTCTGAGGC
	APT-r	CCCGACAACTTCTCACGACCC
qRT-PCR analysis	qAPTf	GGAGCTGCCATCAAATTGCTAGAC
	qAPTf	CCCGACAACTTCTCACGACCC
	qTnt1f	CAGTGCTACCTCCTCTGGATG
	qTnt1r	GGCTACCAACCAAACCAAGTC
Constructs production	oJCBC4	CCGAATTCCCATGGAGTCAAAGATTC
	oJCBC5	CCGAATTCATGGATCGATGTTCGACGTACGTTC
	oJCBC6	CAGCGGCCGCGTCGGCATGCATTCAAACTAG
	oJCBC7	CGCTCGAGTAACGCGAGTAGAAGTTGTTG
	oJCMN13	GTCTCCGAAGTGCCAATGGAG
	oJCMN14	CACCTCCATTGGCACTTCGGAG
	oJCMN15	GGCACAAAAGAATGGGTCATATG
	oJCMN16	CCAACTGCTCCACTTCAAGATC

This fragment was cloned into the PCR8/GW/TOPO cloning vector (Invitrogen) giving the pJCMN4 plasmid. The *Nhe*I-*Nde*I fragment from pJCMN4 was cloned into the *Nhe*I-*Nde*I sites of pJCMN5 to produce the pJCMN7 plasmid. The *BamH*I-*Nco*I fragment from pJCMN7 was then transferred into the corresponding *BamH*I-*Nco*I sites of pBC6 to give the pBC10 (mutated proteins) plasmid. These plasmids were digested with *XhoI* and *NruI* and the Tnt1 protein encoding cassette was cloned into the *XhoI-NruI* site of the pBZRf vector, that carries a 35S::zeoR cassette (from the p35S-loxP-Zeo vector, gift of Pr Hasebe), cloned between two LoxP sites in direct orientation in a pMCS5 backbone (MoBiTec). This resulted in plasmids pBC13 (wt proteins) and pBC14 (mutated integrase). The schematic representation of all plasmids used in this work is shown in Figure 4.1.


A
325 EcoRI (2)
AmpR
pJCBC3
5812 bp
35S-nptII-nos
intron, reverse
2163 EcoRI (2)
attL1
205 EcoRI (2)
ccdB
pENTR3C
2723 bp
632 EcoRI (2)
attL2
KanR
digestion with EcoRI
186 SalI (1)
193 BamHI (1)
pCV1
4139 bp
intron, reverse
35S-nptII-nos
KanR
2054 NotI (1)
2063 XhoI (1)
3' fragment
PCR introducing NotI and XhoI to be cloned
5' fragment
F1 ori
LacZ alpha
689 BamHI (2)
734 SalI (1)
AmpR
5' LTR
short
long
ColE1 origin
1613 BglII (4)
pBSX1
8507 bp
LacO
oJCBC6
Tnt1-coding-proteins
2462 BglII (4)
3' LTR
1D of DDE integrase
Tnt1-coding-proteins
oJCBC7
2D of DDE integrase
E of DDE integrase
4619 BglII (4)
Tnt1-coding-proteins
4280 BglII (4)
3605 BamHI (2)
Long mini-Tnt1
attL1
192 XhoI (2)
5' LTR
KanR
pBC5
7696 bp
Tnt1-coding-proteins
attL2
5619 XhoI (2)
3' LTR
Tnt1-coding-proteins
35S-nptII-nos
intron, reverse
Short mini-Tnt1
attL1
192 XhoI (2)
5' LTR
Tnt1-coding-proteins
pBC7
5704 bp
KanR
intron, reverse
35S-nptII-nos
attL2
3627 XhoI (2)
3'LTR
Tnt1-coding-proteins
276 XhoI (1)
pBHRf
4943 bp
Hygromicin R
AmpR
digestion with XhoI
5' LTR
AmpR
Tnt1-coding-proteins
pBC12
10370 bp
Hygromicin R
intron, reverse
35S-nptII-nos
3' LTR
Tnt1-coding-proteins
5' LTR
Tnt1-coding-proteins
AmpR
pBC11
8378 bp
intron, reverse
35S-nptII-nos
Hygromicin R
Tnt1-coding-proteins
3'LTR


B
AmpR
695 SmaI (2)
734 SalI (1)
5' LTR
pJCMN5
8069 bp
5857 ApaI (2)
5848 SmaI (2)
nos terminator
Tnt1-94
4652 ApaI (2)
attL1
186 SalI (1)
408 SmaI (1)
ccdB
pENTR3C
2723 bp
attL2
KanR
digestion with SalI and SmaI
attL1
483 XhoI (1)
5' LTR
KanR
5902 NruI (1)
attL2
pBC6
7371 bp
nos terminator
1D of DDE integrase
Tnt1-coding-proteins
2D of DDE integrase
E of DDE integrase
270 NruI (1)
276 XhoI (1)
35S
pBZRf
5097 bp
AmpR
ZeoR
digestion with XhoI and NruI
nos terminator
AmpR
pBC13
10253 bp
E of DDE integrase
Tnt1-coding-proteins
2D of DDE integrase
1D of DDE integrase
ZeoR
35S
5' LTR

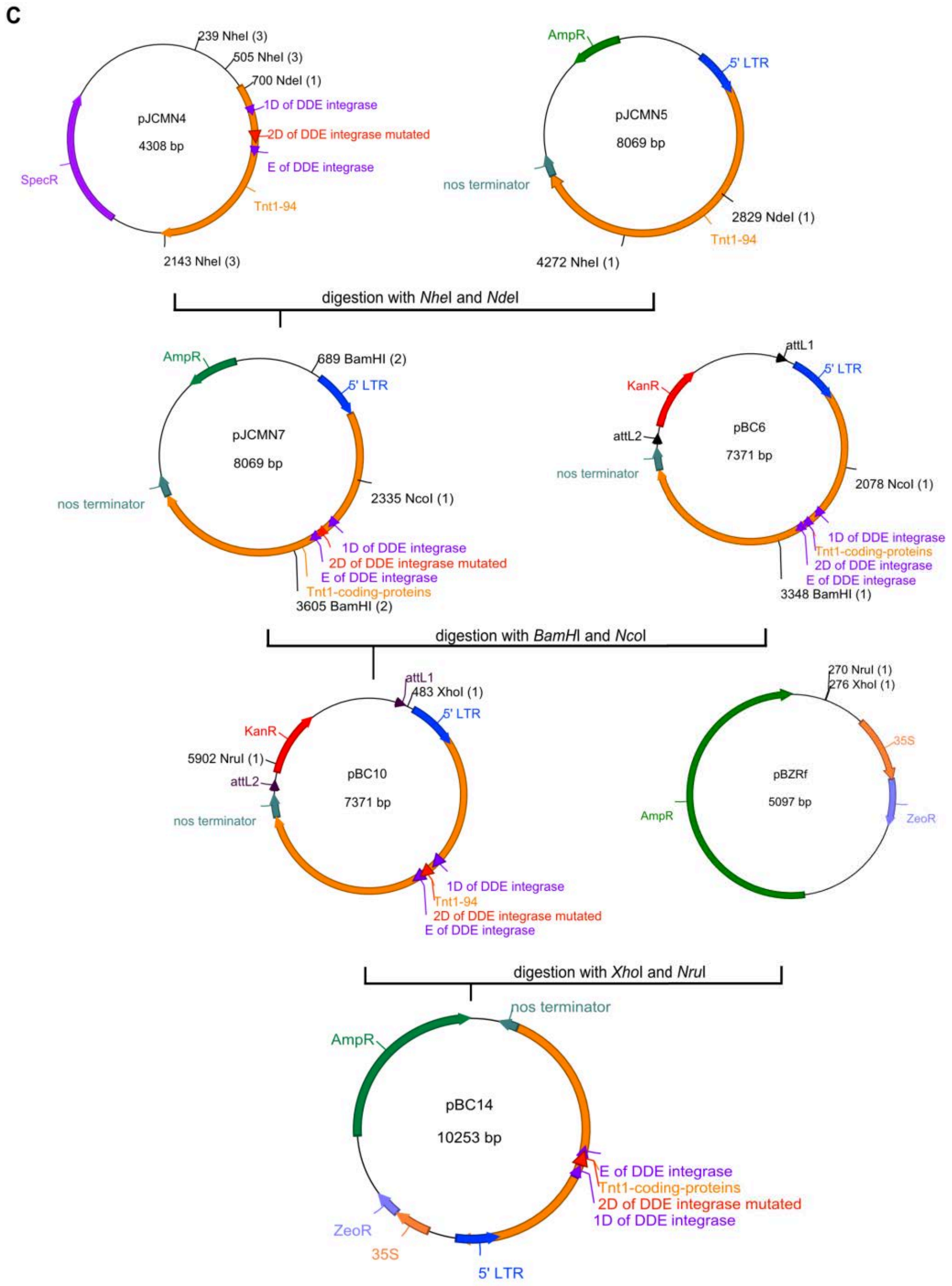


Figure 4 1. Cloning strategy to obtain the plasmids used for the mini-Tnt1 two-component transposition system. (A) Schemes of the cloning strategies used to obtain the tagged mini-Tnt1 elements. Each of the plasmids, digestions and PCR used are shown. (B) Schemes of the cloning strategies used to obtain the plasmid expressing the proteins needed for Tnt1 transposition, and (C) schemes of the cloning strategies used to obtain the plasmid expressing the Tnt1 proteins including a mutated version of the integrase.

Plant transformation and selection

Transformation experiments were performed by protoplast PEG fusion as previously described (Trouiller et al. 2006). A total of $4x10^6$ protoplasts were transformed with supercoiled DNA of plasmid Tnk23. Aliquots of 10 µg of DNA were used to transform $4x10^5$ protoplasts. Protoplasts were plated on cellophane-covered regeneration plates (10^5 protoplasts/plates) containing BCDAT medium with mannitol and incubated in light (15 W/m^2) for 6 days. Antibiotic-resistant plants were selected by transfer of the cellophane overlays for 3 days on BCDAT medium containing hygromycin (Duchefa) (20µg/ml) and zeomycin (20µg/ml) (Duchefa) when appropriate for mini-Tnt1 transformants. The cellophane overlays were transferred to BCDAT medium containing G418 (50 µg/ml) for 10 days.

PCR analysis of transformants

Moss DNA from mini-Tnt1 transformants was prepared as previously described (Trouiller et al. 2006). The primers oBC1 and oBC2 were used to amplify almost the entire mini-Tnt1 element. The primers JCMN23 and JCMN24 were used to amplify the flanking region of intron in reverse orientation with respect to *nptII* of the mini-Tnt1 constructs. The primers oCV1 and oCV2 were used to amplify a region of the Hygromycin gene of the mini-Tnt1 plasmids. The primers JCMN15 and JCMN16 were used to amplify a region of the Tnt1 proteins plasmids. The primers APTg-f, APT-r were used as control. The PCR protocol was: 5 min at 95 °C, 30 cycles of 40 s at 95 °C, 40 s at 60 °C, 1 min at 72 °C, 4 min at 72 °C, and storage at 4 °C. The amplification product of intron in reverse orientation respect to *nptII* was cloned in TOPO TA Cloning^{TM} kit (Invitrogen, Carlsbad, CA, USA) and transformed into *Escherichia coli* strain DH5α by heat shock (Sambrook et al. 1989). Selected clones were grown up and their plasmid DNA was extracted using Wizard Plus SV Minipreps DNA Purification System^{TM} (Promega, Madison, WI, USA). Clones containing the insert were selected by digestion using *EcoR*I and were sequenced using the universal M13 forward and reverse primers. Sequences of primers used in this study can be found in Table 4.1.

Real-time RT-PCR

Total RNA was extracted from 7-day-old cultivated protonema using PureLink^{TM} RNA Mini Kit (Applied Biosystems, Ambion). Genomic DNA was eliminated by treatment with DNA-free^{TM} kit (Applied Biosystems, Ambion). One microgram of total RNA was used to synthesize first-strand cDNA using SuperScript^{TM} III Reverse Transcriptase kit (Invitrogen).

The quantitative real-time RT-PCR reactions (qRT-PCR) were performed on optical 96-well plates in the Roche LightCycler 480 instrument using SYBR Green I Master (Roche Applied Science), primers at 10µM and 1/5 of the cDNA obtained from the reverse transcription of 100 ng of RNA, running each sample in triplicates. The cycling conditions were: 95ºC for 5 minutes (holding stage); then 95ºC for 10 seconds, 56ºC for 10 seconds and 72ºC for 10 seconds (amplification stage); and finally, the qRT-PCR specificity was checked with the melting curve. Reverse transcriptase negative controls and non-template controls were included. The adenine phosphoribosyl transferase (*APT*) gene (Schaefer et al. 2010) was used as internal control to normalize the qRT-PCR output, where Ct values of 40 or above were considered negative values or lack of amplification. Primers were designed using Primer3 software tool (Untergasser et al. 2012). The primers qAPTf and qAPTr were used to amplify the *APT* transcripts, and the primers qTnt1f and qTnt1r were used to amplify Tnt1 transcripts. Sequences of primers used in this study can be found in Table 4.1.

4.4.- RESULTS

Analysis of the expression of Tnt1 elements inserted in the *P. patens* genome

This project started when the group of Dr. Fabien Nogué sent us a number of *P. patens* clones containing insertions of a full Tnt1 element. So, the first objective of this work was to analyze whether the inserted Tnt1 copies in the genome are expressed, which may indicate that they can maintain the potential to transpose again. In case Tnt1 is still able to transpose, the transformed clones will not be stable. We have chosen protonema tissue (filamentous stage) and protoplasts, because Tnt1 in tobacco plant is not expressed in non-stressed tissues and its expression can only be detected in protoplasts or in other stress situations (Grandbastien et al. 2005).

We analyzed the expression of two independent Tnt1-containing clones in protonema tissue, protoplasts and regenerated protonema obtained by cultivating these protoplasts. Figure 4.2 shows that Tnt1 elements are expressed in protonema tissue and there is a repression of the expression in the protoplasts. When cultivated protoplasts have regenerated till obtain protonema again, the expression is recovered demonstrating that the decrease of expression is associated to protoplasts production. This result suggests that the pattern of expression of Tnt1 in *P. patens* is exactly the opposite of Tnt1 in its natural host and in the heterologous plant species where it has been introduced.

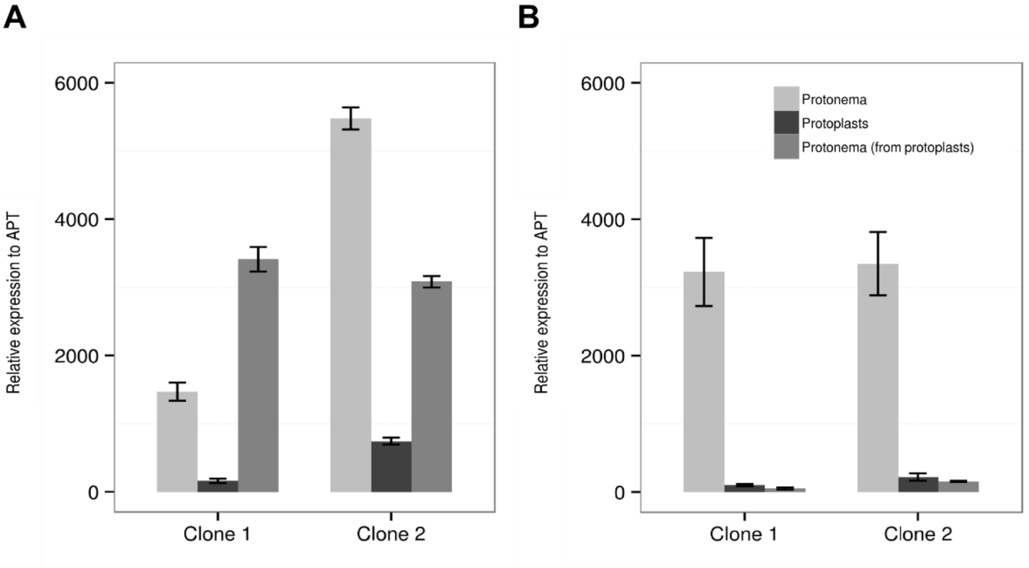


Figure 4.2. Expression analysis of Tnt1(A) and RLG1(B) retrotransposon family in different cell types of two *P. patens* clones transformed with Tnk23 plasmid. Error bars represent +/- SE of three technical replicates.

This peculiar transcription pattern of Tnt1 in *P. patens* may be related to the way *P. patens* regulates its own retrotransposons or to the way it regulates stress-related responses. We therefore decided to compare the expression of the newly introduced Tnt1 with that of endogenous *P. patens* retrotransposons. As shown in Chapter 2, different retrotransposon families are expressed in protonema tissue in *P. patens* (Rensing et al. submitted). So, we analyzed the expression of the largest retrotransposon family, the *Gypsy* RLG1 family which occupies a quarter of the genome of *P. patens*. The results show that RLG1 family is expressed in protonema, whereas its expression decreases in protoplasts (Figure 4.2). Both RLG1 endogenous elements and the tobacco Tnt1 in *P. patens* seem to be regulated in a similar way. Nevertheless, more analyses should be done in order to understand how *P. patens* deal with transposon regulation.

Analysis of a possible increase of Tnt1 copy number over time in *P. patens*

Irrespective of the reasons behind the fact that Tnt1 is transcribed in protonema tissues, this suggests that inserted Tnt1 may have the potential to transpose again. So, we decided to determine if the copy number of Tnt1 that have been inserted into the *P. patens* genome could change over time. A total of 9 clones transformed with Tnk23 plasmid were analyzed by qPCR in two different moments, at the beginning of its culture (t0) and after three months of culture (t3). The results obtained from qPCR were compared with the SSAP analysis of the same clones performed by the group of Dr. Marie-Angèle Grandbastien (INRA AgroParisTech – IJPB, France) (Table 4.2).

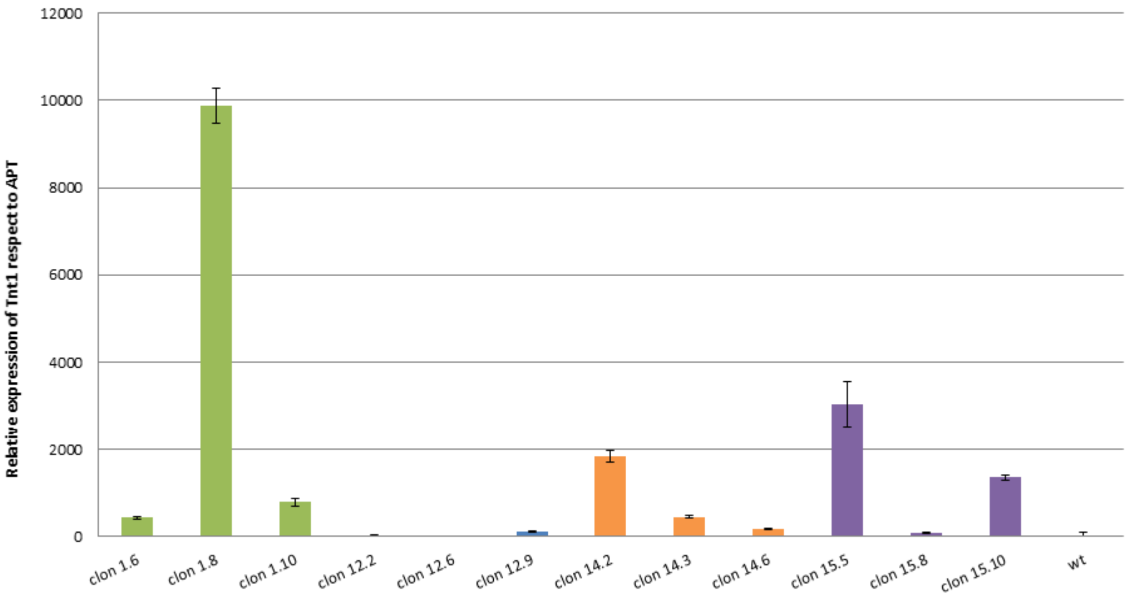
Surprisingly, SSAP detected a higher copy number of insertions than the qPCR estimates for all analyzed clones. We suspected that the clones we analyzed could be heterogeneous (chimeric) in terms of the Tnt1 insertions, which would explain that the number of different insertions in the population could be higher than the mean number of insertions per cell. On the other hand, the overall estimated copy number of these chimeric clones is low (between 1 and 5 copies) before and after 3 months (Table 4.2), showing no major changes with time.

Table 4.2. Comparison of the SSAP results with the quantification of Tnt1 copy number in clones transformed with Tnk23 plasmid. The quantification of Tnt1 copies was performed by qPCR twice, the first one (t0) and the other time after 3 months (t3). As a control, protonema wt tissue was used and n.d. stands for no data.

Clone number	SSAP quantification	qPCR t0	qPCR t3
4	27	1.92	1.62
7	29	2.37	2.16
11	40	1.83	1.68
25	24	5.24	4.88
27	n.d.	5.02	4.34
28	6	2.1	2.18
29	12	2.09	1.96
31	17	1.47	2.25
39	n.d.	2.14	2.11
wt	0	0.55	0.55

Due to the heterogeneity of the material, those results may not be conclusive, and we decided, in collaboration with Dr. Fabien Nogué, to transform and obtain a new small set of transformed clones. The idea was to transform again with Tnk23 plasmid and select 4 clones that contain inserted Tnt1.

We consider this material as potentially chimeric. Protoplasts were isolated from four chimeric original clones (clones 1, 12, 14 and 15) and clones derived from a single protoplast, and therefore homogeneous, were obtained. We analyzed the expression level in protonema, as well as the Tnt1 copy number of these 12 homogeneous clones. Figure 4.3 shows that the copy number of the homogeneous clones is different. For instance, the 3 coming from clone 1 have similar copy number (between 2 and 3 copies), the 3 from clone 12 have very low copy number (two of them may have no Tnt1 insertions) and homogenous clones from 14 and 15 clones vary between 1 and 4 copies. This may suggest that, indeed, the original clones were heterogeneous. Regarding the expression, we observed that there is a huge variability among clones (Figure 4.3). Part of this may be explained by the copy number, for example absence of expression for 12.2 and 12.6 clones, but other factors such as insertion locus or just stochasticity may also contribute.



Num. Copies Tnt1													
t0	2.64	2.59	2.45	0.47	0.56	1.42	3.21	2.42	1.53	2.17	2.40	1.80	
t8	1.63	1.59	1.42	1.19	0.20	1.02							

Figure 4.3. Expression analysis of Tnt1 and quantification of the Tnt1 copy number in the 12 homogeneous clones. Error bars represent +/- SE of three technical replicates. In the lower part, the estimated copy number of Tnt1 by qPCR is indicated per each clone, the first one (t0) and the other time after 8 months (t8).

We estimated again the Tnt1 copy number by qPCR after 8 months of culture (t8) of the homogenous clones from 1 and 12 chimeric original clones (Figure 4.3). Although no major differences over time were found, the integrated Tnt1 are able to transcribe and may be able to transpose increasing the number of copies.

More analyses are required to determine under which conditions the integrated Tnt1 can be mobilized within the genome. However, the expression pattern and the potential transposition of Tnt1 could be a problem for using Tnt1 to generate mutants in *P. patens*.

Design of a Tnt1-based two-component system for insertion mutagenesis in *P. patens*

Although we have not been able to detect a significant increase in Tnt1 copy number under laboratory conditions, the expression of Tnt1 in non-stressed protonema tissue suggests that the integrated Tnt1 elements may be able to transpose, which can be seen as a problem for using Tnt1 to produce a mutant collection. We therefore decided to design a Tnt1-based two-component vector system, where the mobile unit is separated from the sequences needed to express the proteins required for its retrotransposition. The idea is to generate a defective element that can be mobilized in *trans* by the transient expression of required proteins. With the help from Beatriz Contreras, master student from our lab, we constructed different mini-Tnt1 elements, replacing a variable fraction of coding sequence by a selective marker obtaining the long mini-Tnt1 element (pBC12) and the short mini-Tnt1 element (pBC11) (Figure 4.4).

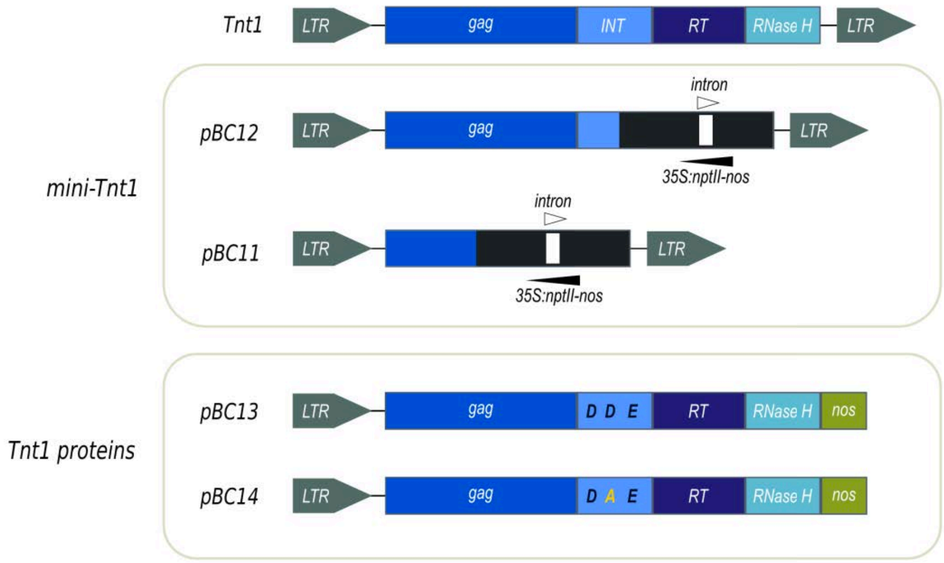


Figure 4.4. Tnt1-based two-component retrotransposon system**.** The LTRs (LTR) of the Tnt1 element are shown in grey and the different proteins it encodes are shown in boxes filled with different shades of blue: gag (gag), integrase (INT), reverse transcriptase (RT) and RNase H (RNase H), whereas *nos* terminator is in green.

The marker allows selecting when transposition events have occurred, as it has been reported before (Hou et al. 2010). The antibiotic resistance gene *nptII* is interrupted by an intron in reverse orientation with respect to its promoter, but the direct orientation with respect to the Tnt1 promoter. So, this mechanism makes sure that resistance is achieved only after transcription and retrotransposition of the Tnt1 defective element (Figure 4.4). The plasmid used to express the proteins required for transposition was constructed using the Tnt1 plasmid where the 3'LTR of Tnt1 was replaced by a *nos* terminator which makes it unable to retrotranspose as the two LTRs are required for reverse transcription (Figure 4.4). In addition, as a control, we also obtained a construct expressing a defective version of the Tnt1 proteins by introducing a mutation in the integrase core domain that blocks integration (Ke and Voytas 1999) (Figure 4.4).

We transformed *P. patens* with different plasmid combination in order to assess the transposition of mini-Tnt1. *P. patens* clones resistant to kanamycin were only obtained with the long and the short version of mini-Tnt1 (pBC12) together with plasmid expressing the wild type Tnt1 proteins (pBC13) (Table 4.3).

Table 4.3. Co-transformation results of mini-Tnt1 and protein constructs in *P. patens*

Mini-Tnt1 constructs	**Protein constructs**	**Num of transformed protoplasts**	**Num of KanR clones**
pBC12 (long)	pBC13 (wt)	2108	34
pBC12 (long)	pBC14 (mutated)	1576	0
pBC12 (long)	None	3292	0
pBC11 (short)	pBC13 (wt)	2304	2

No resistant clones were obtained when we combined the mini-Tnt1 plasmid with the mutated version of the Tnt1 proteins or when using only the mini-Tnt1 plasmid (Table 4.3). These results suggest that retrotransposition has occurred and mini-Tnt1 elements have been inserted into the genome through that mechanism.

Interestingly, more clones were obtained with the longer version of the mini-Tnt1 (34 clones) than with the shorter version (2 clones) (Table 4.3), which suggest that some internal Tnt1 sequence may be required for high efficient transposition.

The resistant kanamycin clones were analyzed by PCR, where the expected size of mini-Tnt1 sequence can be amplified in all samples (Figure 4.5). Also, PCR and sequencing results show that the *nptII* gene intron was spliced-out in all cases, confirm that mini-Tnt1 elements have been inserted into the genome by retrotransposition mechanism (Figure 4.5).

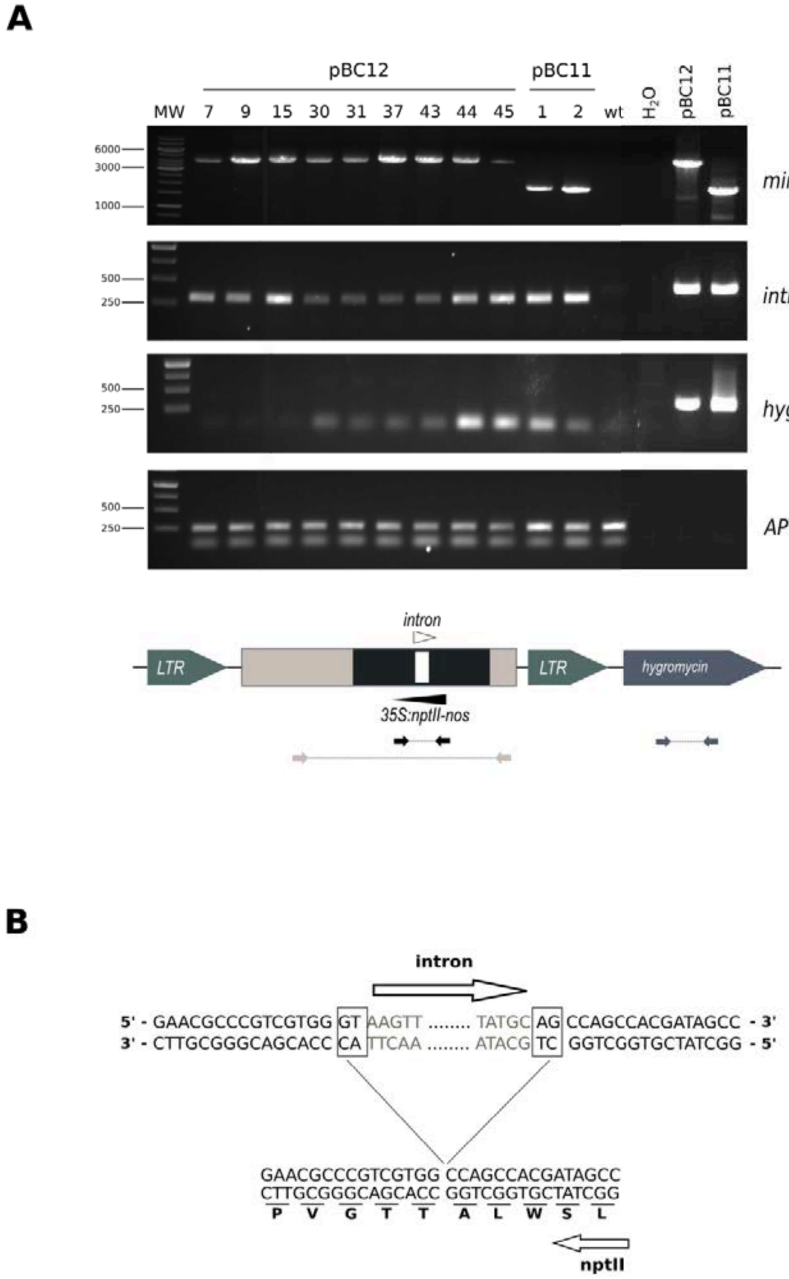

A
pBC12
pBC11
MW 7 9 15 30 31 37 43 44 45 1 2 wt H_2O pBC12 pBC11
6000
3000
1000
mini-Tnt1
500
250
intron
500
250
hygromycin
500
250
APT
intron
LTR
LTR
hygromycin
35S:nptII-nos
B
intron
5' - GAACGCCCGTCGTGG GT AAGTT .... TATGC AG CCAGCCACGATAGCC - 3'
3' - CTTGCGGGCAGCACC CA TTCAA .... ATACG TC GGTCGGTGCTATCGG - 5'
GAACGCCCGTCGTGG CCAGCCACGATAGCC
CTTGCGGGCAGCACC GGTCGGTGCTATCGG
P V G T T A L W S L
nptII

Figure 4.5. Analysis of the presence of mini-Tnt1 and other regions of the construct. A) 10 transformed *P. patens* clones have been analyzed, 2 of them being transformed with pBC11 (short mini-Tnt1 construct) and 8 with pBC12 (long mini-Tnt1 construct). B) Splicing of the intron sequence. The band amplified with *nptII* primers flanking the intron was cloned and sequenced. The sequence of the amplified product is shown below the sequence of the plasmid *nptII* gene, showing that the intron sequence was correctly spliced.

The group of Dr. Marie-Angèle Grandbastien (INRA AgroParisTech – IJPB, France) performed an SSAP analysis on 14 clones. Figure 4.6 shows that the copy number varies from one to two copies per clone.

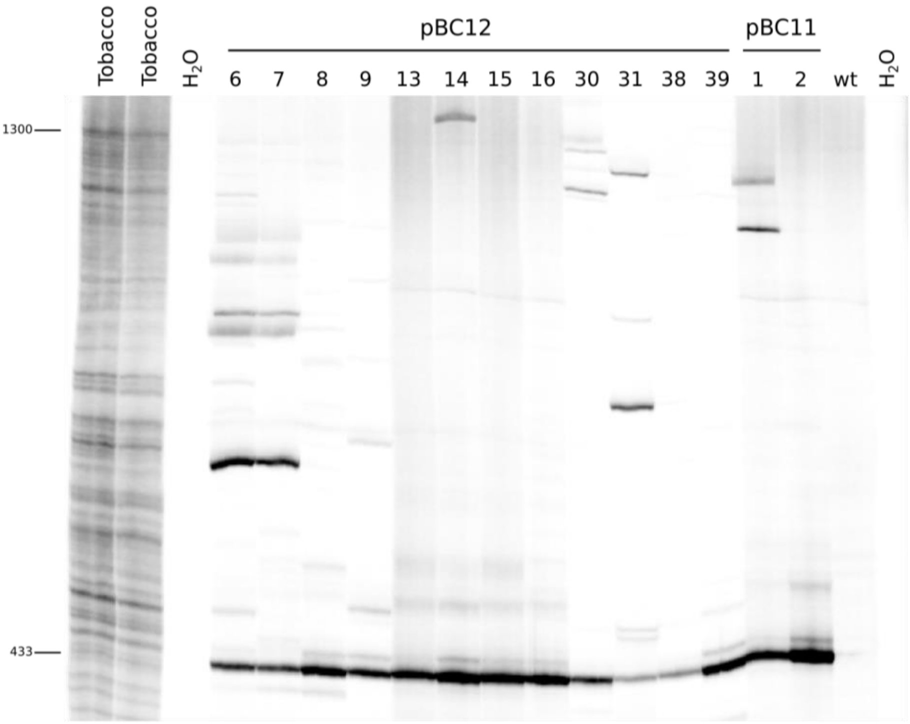


Figure 4.6. SSAP analysis of mini-Tnt1 insertions in *P. patens* clones either transformed with pBC11 (short mini-Tnt1 construct – 12 clones) or with pBC12 (long mini-Tnt1 construct – 2 clones) together with the pBC13 plasmid harboring the Tnt1wild type proteins necessary for the retrotransposition process achievement.

A total of four bands were cloned and sequenced, determining that Tnt1 sequences start with the first nucleotide of the 5' LTR in all cases, and no plasmid sequence is inserted. This result confirms that the mini-Tnt1 elements are integrated into the genome through retrotransposition mechanism.

These sequences also allowed to determine the insertion site of the mini-Tnt1 elements revealed that mini-Tnt1 elements inserted within (3) or at less than 100 nt (1) of annotated *P. patens* genes (Table 4.4).

Table 4.4. Analysis of sequences flanking mini-Tnt1 insertions. The mini-Tnt1 constructs correspond to pBC12 (long) and pBC11 (short).

Tnt1 or mini-Tnt1 insertion	**Chromosome #**	**Tnt1 position**	**Closest gene (Phytozome and Cosmoss numbers)**	**Coordinates of closest gene (ATG to stop)**	**Distance to ATG or stop**	**JGI annotation of closest gene**
pBC12-6	Chr_23	7136546	Pp3c23_10200 Pp1s10_17V6.1	7137132-7134014	Between ATG and STOP Exon1	Auxin efflux carrier component 3-related
pBC11-1	Chr_7	10718730	Pp3c7_15700 Pp1s153_79V6.1	10717393-10719080	Between ATG and STOP Exon4	F-box-like
pBC12-30	Chr_18	7691183	Pp3c18_10870 Pp1s19_291V6.1	7688119-7695553	Between ATG and STOP Exon11	Protein tyrosine kinase
pBC12-14	Chr_9	3889027	Pp3c9_6830 Pp1s220_62V6.1	3889604-3892995	577 pb from ATG	no functional annotations

All these results show that Tnt1-based two-component vector system transposes efficiently and allow us to select for clones in which mini-Tnt1 has been inserted into the genome of *P. patens*. These results also show that strong the insertion preference into genes shown for Tnt1 is conserved in the mini-Tnt1 elements. All these results confirm that the mini-Tnt1 system is an ideal tool to be used for insertional mutagenesis in *P. patens*.

4.5.- DISCUSSION

The high efficiency of homologous recombination in *P. patens* makes it an ideal system for reverse genetic analyses. On the contrary, its inefficiency of *P. patens* for integrating DNA with no sequence similarity makes it very difficult to generate insertion mutants for forward genetic analyses in this species. Previous work in the laboratory allowed the development of a highly efficient system to create insertional mutants in *P. patens* based on the transposition of the tobacco Tnt1 retrotransposon.

However, the expression analysis showed that the integrated Tnt1 elements are expressed in protonema tissue (filamentous stage) and their expression decreases in protoplasts. This was unexpected based on the previous knowledge on Tnt1 expression in its host tobacco and in heterologous species. This opened a fundamental question on the expression of Tnt1, and in general on TEs in *P. patens*, but also suggested important limitations for the use of Tnt1 as a tool for insertional mutagenesis in *P. patens*.

The analysis of Tnt1 expression showed that after a decrease in expression in protoplasts, when those protoplasts were cultivated to develop new protonema tissue, Tnt1 expression was recovered, confirming that Tnt1 is expressed in non-stressed protonema tissue and that its expression is transiently inhibited when producing protoplasts. The Tnt1 expression in tobacco and in other plants where Tnt1 has been introduced is very low in non-stressed vegetative tissues and is induced by stress and by protoplasts isolation (Beguiristain et al. 2001; Grandbastein et al. 2005; D'Erfurth et al. 2003; Mazier et al. 2007; Tadege et al. 2008; Cui et al. 2013). The pattern of expression observed when this element is introduced in *P. patens* is, therefore, the opposite of the one reported in other organisms.

The transcriptional activation of Tnt1 in tobacco and in heterologous hosts is linked to the activation of plant stress responses, which seems to be a common feature of different plant TEs (Grandbastien et al. 2005). Therefore, a different pattern of expression in *P. patens* could be due to a different regulation of stress responses or a different regulation of TEs in this species. Finally, it could also be a difference restricted to the transcription factors inducing Tnt1 in this species. Although *P. patens* presents some particularities in its stress responses, most plant responses seem to be already present in this evolutionary basal species (Ponce de León et al. 2012). In order to determine if the atypical pattern of Tnt1 expression in *P. patens* is specific for this element, we analyzed the expression of the endogenous retrotransposon RLG1 family in the same tissues. This RLG1 family, which by itself accounts for a quarter of the *P. patens* genome, shows the same pattern of expression as Tnt1. RLG1 family shows a high transcription in protonema and a decrease in protoplasts. This suggests that retrotransposons are regulated differently in *P. patens* and that the differences shown for Tnt1 regulation reveal more profound differences of the way *P. patens* deals with stress and transposon regulation as compared to most plants.

The pattern of expression of Tnt1 in *P. patens*, with expression in non-stressed protonema cells, suggests that after integration Tnt1 may continue to transpose in cultured protonema. This fact may lead to chimeric populations of cells within protonema tissue. We performed experiments trying to correlate the expression level with the number of Tnt1 copies integrated in the genome and performed qPCR analyses to determine Tnt1 copy number. The number of the different copies of Tnt1 determined by the number of SSAP bands for the analyzed clones was higher than the mean copy number obtained from qPCR. This suggests that Tnt1 transposed after integration obtaining a chimeric material with respect to Tnt1. This would explain the relatively high copy number of insertions detected by SSAP and the low copy number calculated by qPCR.

Due to the heterogeneity of the material, we analyzed the expression level as well as the Tnt1 copy number in homogeneous clones. The fact that the copy number of the homogeneous clones is different indicates that the original clones were heterogeneous. Although no major changes of the copy number were estimated in both chimeric and homogeneous clones over time, the expression and potential transposition of Tnt1 in protonema may generate unstable mutants. More experiments should be performed to determine in which conditions Tnt1 is able to transpose in *P. patens*.

The first goal of this project was to develop an efficient tool for insertion mutagenesis in *P. patens*. The atypical pattern of expression of Tnt1 in this organism, and in particular its expression in non-stressed protonema tissue, makes newly inserted Tnt1 elements potentially able to continue to generate new mutations and, therefore, could make the obtained phenotypes unstable. This leads us to develop a Tnt1-based two-component vector system where the mobile Tnt1 unit would be stabilized.

We have constructed a mini-Tnt1 element which contains a retrotransposition indicator selectable gene replacing the original coding region. This element cannot transpose unless it is activated by a vector expressing Tnt1 proteins in *trans*. The results show that of the two versions of the mini-Tnt1 used, the long mini-Tnt1 performs better, suggesting that a certain length of internal Tnt1 sequence may be required for efficient retrotransposition and integration into the genome.

Our results also show that the expression of wild type Tnt1 proteins is required for Tnt1 transposition, which stresses the need of an active integrase provided in *trans* for mini-Tnt1 transposition and confirms that the mini-Tnt1 elements cannot use the retrotransposition machinery from the endogenous TEs of *P. patens*. As the Tnt1 sequences encoding for Tnt1 proteins are not integrated into the genome, once the mini-Tnt1 is integrated in the genome, this unit will not be able to transpose again. In this way, the mini-Tnt1 insertions are stabilized obtaining potential phenotypes to be studied.

Previous analyses in the laboratory showed that Tnt1 targets genic regions for integration in *P. patens*. Indeed, the analysis of 22 independent Tnt1 insertion sites showed that in 54% of the cases Tnt1 was inserted within a known protein coding gene, while these sequences only represent the 17% of *P. patens* genome (Rensing et al. 2008).

Interestingly, the all four mini-Tnt1 insertions analyzed are located less than 1 kb from a gene, with three of them being located within a gene. This result indicates that mini-Tnt1 has maintained the clear preference for integrating into genic regions shown for the complete Tnt1 element. Other *Copia*-like retrotransposon as Tto1 or Tos17 present the same preference inserting into euchromatic regions (Okamoto and Hirochika 2000; Miyao et al. 2003), and also Tnt1 in its natural host (Le et al. 2007). These Tnt1 insertions may potentially alter gene regulation or result in epigenetic gene silencing. More experiments should be done to study expression pattern of these genes which can lead to diverse phenotypes.

GENERAL DISCUSSION

5.- GENERAL DISCUSSION

Transposons are mutagenic elements and can be an important source of genetic variability. Their impact and significance depend on the perspective at which we study them, and sometimes may seem contradictory. On one hand, TEs are invasive and are expanded in bursts of transposition compromising host viability. On the other hand, although most TE insertions are neutral or deleterious and are selected against and lost, few TE insertions can confer a selective advantage. The work presented in this dissertation provides a wide view on the different TE impacts on host genomes, from genome-wide scale analysis to the study of impact of particular TE insertions. Moreover, this work also presents a study in which we have used TEs as a genetic tool for obtaining mutant collections.

At a genome-wide level, we have investigated how TEs have shaped melon and cucumber genomes. This study has shown that TEs can modify the structure of chromosomes and the landscape of TEs. And this fact may impact the evolution of genes located in different TE-defined regions.

The recent TE activity in melon after the melon-cucumber split (10 Mya, Sebastian et al. 2010) seems to be the responsible for the difference in genome size between these two species. In addition, whole-genome analyses of these and other related cucumber species established that the ancestor of these group of *Cucumis* species had 12 chromosomes, which have been maintained in melon but not in cucumber. Indeed, the seven cucumber chromosomes arose from fusions and rearrangements of the 12 ancestral chromosomes (Huang et al. 2009; Li et al. 2011a; Garcia-Mas et al. 2012).

The comparison of the TE chromosomal distribution in melon and cucumber shows that, although TEs seem to accumulate in the pericentromeric regions of the chromosomes in both species, the extent of this accumulation and the size of the pericentromeric regions are very different. Indeed, our results show that the recent TE activity in melon has expanded the chromosomal pericentromeric regions.

Interestingly, the large melon pericentromeric regions show a clear reduction of the recombination rate, whereas the small and less TE-rich pericentromeric regions in cucumber do not show this effect, and the recombination frequency is essentially constant along the chromosomes in this species. This effect has also been recently seen for *Arabis alpina* where TEs have expanded the pericentromeric regions as compared to *A. thaliana* and *A. lyrata*, extending also the low recombination region, which in *A. thaliana* is restricted to the centromere, to the large pericentromeric regions in this species (Willing et al. 2016).

As a consequence of this increase of the pericentromeric regions, genes that were located in gene-rich highly recombining regions of the chromosome may be now located in TE-rich low recombining regions of the melon genome. Taking advantage of the high collinearity of chromosomes one of melon and seven of cucumber, we have obtained a list of orthologous genes and analyzed their location in melon and cucumber. Although most genes are located in similar regions, 188 melon genes are located in a TE-rich low recombining regions whereas in cucumber they sit in a gene-rich region. This suggests that the expansion of the melon pericentromeric regions may have had an impact on the evolution of some melon genes.

Currently, this analysis is being pursued in collaboration with the groups of Drs. Jordi Garcia-Mas and Sebastián Ramos-Onsins, and the results obtained so far suggest that the melon TE-rich low recombining regions concentrate melon specific genes, whereas genes orthologous to cucumber or present in other organisms are almost exclusively found in gene-rich regions. These results suggest that TE accumulation, and consequently the reduction of the recombination, may allow some genes to explore new diversity and evolve new functions. Although we couldn’t determine a functional enrichment in this small subset, it is an interesting approach to understand how TEs affect in the evolution of genes in different genomic compartments.

This analysis has provided information about the TE activity in the evolution of these two *Cucumis* species. But, the analysis of polymorphisms due to TEs in a wide range of varieties of the same species can also provide information about the recent TE activity.

Since the discovery of TEs, the studied cases were those in which the TE insertions conferred detectable phenotypes. But the new bioinformatic tools, as the ones used in this thesis, can identify genome-wide TE movements by comparing different varieties to a reference genome.

In this work, we have investigated the recent activity of TEs in three distinct species: melon, date palm and *P. patens*. We performed the analysis of TE insertion polymorphisms in three distinct species, having different quality of the reference genome assembly, the available resequencing data and the reference genome TE annotation. After performing these analyses, we can conclude that the quality of genome assembly is the most important requirement allowing to obtain a correct annotation of the TE fraction and to perform a proper analysis of TE insertion polymorphisms. However, a high enough coverage of the resequencing data is also key to allow polymorphic TE insertions to be detected.

As a general trend, most TE insertion polymorphisms in all the species we analyzed are due to LTR retrotransposons which account for a major TE fraction in the three genomes, like in other plant genomes (Lisch 2012). However, the three studies present also differences. In the melon study, a small number of TE families are responsible for the majority of polymorphisms detected among the seven melon varieties. But in the case of *P. patens*, the vast majority of polymorphisms are due to a specific *Gypsy* family, RLG1, consistent with the fact that this family represents a quarter of the moss genome.

As outlined in the introduction, TE insertions are generally deleterious and the ones inserted within or close to genes are usually selected against (Hollister and Gaut 2009). However, we identified a number of TE-related polymorphisms that are located close or within a gene in the three analyzed genomes, suggesting that some of these insertions provide the genome some advantage. Genes impacted by one of these TE-related polymorphisms could therefore constitute interesting cases to check if they are related to important traits. For instance in melon, a total of 165 PM-TE located in coding regions were identified between the two elite lines PS and VED. The fact that these two lines differ in different traits makes them ideal candidates to assess the importance of TEs in the domestication process. Although the results obtained from selected candidate genes weren’t promising, the possibility that one PM-TE of this list may alter an important agronomical trait is highly likely.

As mentioned in the introduction, when a TE inserts within a gene, it is expected that TE modify the coding capacity of that gene. Among possible modifications, insertion upstream of a gene can modify gene expression by attracting epigenetic silencing machinery or by providing new regulatory elements. This is the case of MITE families which have amplified and redistributed the binding sites of E2F transcription factor during *Brassica* evolution. This study showed that E2F BSs within TEs are bound by the E2F protein *in vivo*, irrespective to their location with respect to genes.

The presence of E2F BS in TEs have two different impacts on the genome on one hand the impact of the E2F-TE amplification and on the other the impact of their relocation. Depending on the location of E2F-TEs respect to genes, they may impact at distinct levels in the regulation of E2F transcriptional network.

For instance, the ones located close to genes may directly incorporate new genes into the E2F transcriptional network. Moreover, some of those E2F-TE insertions close to genes are polymorphic among *A. thaliana* accessions and may induce variability in the E2F regulatory network within the species.

But the vast majority of E2F-TEs are found far from genes. As the E2F-TEs found far from genes are able to bind E2F, they may reduce the E2F proteins available to bind the E2F-TEs in gene promoters. Moreover, the E2F-TEs located far from genes may be mobilized during evolution and they can be a reservoir to rewire new genes into the E2F transcriptional network.

This dissertation has shown that TEs can impact genes and genomes in many different ways, which highlights their importance for the evolution of eukaryote genomes, and more precisely for those of crop plants.

Apart from the interest to study TE dynamics to understand the evolution of eukaryote genomes, the analysis of TEs can also have more applied interests. Indeed, TEs can be used as tools to develop, for example, molecular markers and insertional mutant collections. The last part of this work analyzed the capacity of the tobacco Tnt1 retrotransposon to transpose in the bryophyte species *Physcomitrella patens* with the objective to generate stable insertion mutant collections.

The atypical pattern of expression of Tnt1 in this organism, and in particular its expression in non-stressed protonema tissue, makes newly inserted Tnt1 elements potentially able to continue to generate new mutations and, therefore, could make the obtained phenotypes unstable. This leads us to develop a Tnt1-based two-component vector system where the mobile Tnt1 unit would be stabilized. We have constructed a mini-Tnt1 element which contains a retrotransposition indicator selectable gene replacing the original coding region. This element cannot transpose unless it is activated by a vector expressing Tnt1 proteins in *trans*.

The mini-Tnt1 system overcomes some limitations compared to other insertional mutagens, like T-DNA or *Agrobacterium*. The mini-Tnt1 system can be useful in plants with large genomes because of insertional preference within or close to genes. The mini-Tnt1 vector has a marker gene which facilitates the identification of cell with transposition events and the requirement for the expression of the proteins needed for transposition in *trans* makes the obtained phenotypes stable. This system is particularly suited for *P. patens* where insertion mutants with the conventional approaches are very difficult to obtain.

To sum up, this work has contributed to analyze how TEs impact plant genes and genomes. Different approaches and tools have been used to assess the objectives.

Despite the fact that TEs were called junk DNA, studies, like the ones presented in this dissertation, may allow to understand the role of TEs and now they are considered important elements for genome evolution.

CONCLUSIONS

6.- CONCLUSIONS

1. The transposon content has been annotated in melon and in two cucumber genomes using the REPET package, which has annotated as TE-related sequences up to 43% of melon genome, whereas TEs covers the 26% of both cucumber genomes.

2. The chromosomal distribution shows anti-correlation between TEs and genes in both melon and cucumber genomes, defining two different regions depending on TE-richness. In melon, TEs tend to concentrate in the pericentromeric region of melon chromosomes. Cucumber chromosomes present TE-rich regions quite small and interrupted in some cases by small gene-rich regions.

3. Whereas the euchromatic chromosomal regions of both syntenic chromosomes (chromosome 1 of melon and chromosome 7 of cucumber) span a region of similar length, the pericentromeric region in the melon chromosome are much larger due to the accumulation of more TEs. This suggests that the higher TE activity in melon genome has resulted in an increase of pericentromeric regions.

4. The TE-rich pericentromeric melon regions also contain some genes, which may have evolved differently due to the lower recombination rate of these regions.

5. The polymorphic TE insertions have been detected in the three analyzed genomes: melon, date palm and *Physcomitrella patens*.

6. A total of 2,735 TE insertion polymorphic sites have been identified across the 7 melon varieties. The vast majority of TE insertion polymorphic sites are categorized as retrotransposons (65%) or DNA transposons (32.2%) and only 2.8% of them couldn’t be categorized due to their complex nature.

7. Around 60% of TE insertion polymorphic sites are present in only one melon variety, which indicates that TEs have been actively transposing during the domestication and breeding of melon varieties.

8. In the melon analysis, the 22.3% of 2,735 TE-related polymorphic sites are located in genes, and 13.2% in coding regions.

9. In the 69 date palm varieties, we have identified 117,435 insertions and 7,616 deletions. Two-thirds of these insertions and deletions were caused by retrotransposons, whereas the remaining one-third corresponds to DNA transposons.

10. A total of 1,390 TE insertion polymorphic sites using Main TE annotation and 1,572 TE insertion polymorphic sites using REPET TE annotation have been identified between two *Physcomitrella patens* strains. Over the 90% of the TE insertion polymorphic sites are categorized as retrotransposons.

11. In the *P. patens* analysis, around 10% of TE insertion polymorphic sites are found less than 1kb from a gene, which 37 and 42 of them are located in coding regions, using Main TE annotation and REPET TE annotation, respectively.

12. The E2F-TEs located far from genes are associated with heterochromatic marks, but not for those E2F-TEs located close or within genes.

13. The E2F transcription factor can bind *in vivo* to the E2F binding sites within TEs, regardless the epigenetic mark context, and their distance to genes.

14. The E2F-TEs have been mobilized recently and some of the E2F-TEs located close to genes are polymorphic among *A. thaliana* ecotypes.

15. The tobacco Tnt1 retrotransposon is expressed in non-stressed protonema tissue, which may allow its transposition integrating new Tnt1 elements in the *P. patens* genome.

16. The Tnt1-based two-component vector system has been developed, where the mobile unit is separated from the coding sequences required for its retrotransposition. The defective version of Tnt1, named mini-Tnt1, contains a selection marker only active after retrotransposition.

17. The mini-Tnt1 element transposes efficiently in *P. patens* in the presence of Tnt1 proteins expressed from a different plasmid and therefore, the integrated mini-Tnt1 elements are no longer able to transpose again.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmed, Ikhlak, et al. "Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis." *Nucleic Acids Research* 39.16 (2011): 6919-6931.

Al-Mssallem, Ibrahim S., et al. "Genome sequence of the date palm Phoenix dactylifera L." *Nature communications* 4 (2013).

Alonso, Jose M., et al. "Genome-wide insertional mutagenesis of Arabidopsis thaliana." *Science* 301.5633 (2003): 653-657.

Ambrožová, Kateřina, et al. "Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies." *Annals of botany* (2010): mcq235.

Arumuganathan, Ka, and E. D. Earle. "Nuclear DNA content of some important plant species." *Plant molecular biology reporter* 9.3 (1991): 208-218.

Ashton, N. W., and D. J. Cove. "The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the moss, Physcomitrella patens." *Molecular and General Genetics MGG* 154.1 (1977): 87-95.

Bailey, Jeffrey A., Ge Liu, and Evan E. Eichler. "An Alu transposition model for the origin and expansion of human segmental duplications." *The American Journal of Human Genetics* 73.4 (2003): 823-834.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. "Repbase Update, a database of repetitive elements in eukaryotic genomes." *Mobile DNA* 6.1 (2015): 11.

Beguiristain, Thierry, et al. "Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants." *Plant physiology* 127.1 (2001): 212-221.

Bennetzen, Jeffrey L., Jianxin Ma, and Katrien M. Devos. "Mechanisms of recent genome size variation in flowering plants." *Annals of botany* 95.1 (2005): 127-132.

Bennetzen, Jeffrey L., and Hao Wang. "The contributions of transposable elements to the structure, function, and evolution of plant genomes." *Annual review of plant biology* 65 (2014): 505-530.

Bergman, Casey M., and Hadi Quesneville. "Discovering and detecting transposable elements in genome sequences." *Briefings in bioinformatics* 8.6 (2007): 382-392.

Biswas, Anup K., and David G. Johnson. "Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage." *Cancer research* 72.1 (2012): 13-17.

Bonhomme, Sandrine, et al. "Usefulness of Physcomitrella patens for studying plant organogenesis." *Plant Organogenesis: Methods and Protocols* (2013): 21-43.

Bourque, Guillaume, et al. "Evolution of the mammalian transcription factor binding repertoire via transposable elements." *Genome research* 18.11 (2008): 1752-1762.

Bourque, Guillaume. "Transposable elements in gene regulation and in the evolution of vertebrate genomes." *Current opinion in genetics & development* 19.6 (2009): 607-612.

Bowler, Chris, et al. "Chromatin techniques for plant cells." *The Plant Journal* 39.5 (2004): 776-789.

Bundock, Paul, and Paul Hooykaas. "An Arabidopsis hAT-like transposase is essential for plant development." *Nature* 436.7048 (2005): 282-284.

Butelli, Eugenio, et al. "Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges." *The Plant Cell* 24.3 (2012): 1242-1255.

Casacuberta, Elena, and Josefa González. "The impact of transposable elements in environmental adaptation." *Molecular ecology* 22.6 (2013): 1503-1517.

Casacuberta, Josep M., and Néstor Santiago. "Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes." *Gene* 311 (2003): 1-11.

Cavrak, Vladimir V., et al. "How a retrotransposon exploits the plant's heat stress response for its activation." *PLoS Genet* 10.1 (2014): e1004115.

Choulet, Frédéric, et al. "Structural and functional partitioning of bread wheat chromosome 3B." *Science* 345.6194 (2014): 1249721.

Contreras, Beatriz, et al. "The impact of transposable elements in the evolution of plant genomes: from selfish elements to key players." *Evolutionary biology: Biodiversification from genotype to phenotype*. Springer International Publishing, 2015. 93-105.

Cove, David J., et al. "Culturing the moss Physcomitrella patens." *Cold Spring Harb Protoc* 2009 (2009).

Cowley, Michael, and Rebecca J. Oakey. "Transposable elements re-wire and fine-tune the transcriptome." *PLoS Genet* 9.1 (2013): e1003234.

Cui, Yaya, et al. "Tnt1 retrotransposon mutagenesis: a tool for soybean functional genomics." *Plant physiology* 161.1 (2013): 36-47.

DeGregori, James, and David G. Johnson. "Distinct and overlapping roles for E2F family members in transcription, proliferation and apoptosis." *Current molecular medicine* 6.7 (2006): 739-748.

Denoeud, France, et al. "The coffee genome provides insight into the convergent evolution of caffeine biosynthesis." *Science* 345.6201 (2014): 1181-1184.

d'Erfurth, Isabelle, et al. "Efficient transposition of the Tnt1 tobacco retrotransposon in the model legume Medicago truncatula." *The Plant Journal* 34.1 (2003): 95-106.

Du, Chunguang, et al. "The polychromatic Helitron landscape of the maize genome." *Proceedings of the National Academy of Sciences* 106.47 (2009): 19916-19921.

De Veylder, Lieven, et al. "Control of proliferation, endoreduplication and differentiation by the Arabidopsis E2Fa–DPa transcription factor." *The EMBO journal* 21.6 (2002): 1360-1368.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons." *BMC bioinformatics* 9.1 (2008): 18.

Ewing, Adam D. "Transposable element detection from whole genome sequence data." *Mobile DNA* 6.1 (2015): 1.

Fedoroff, Nina V. "About maize transposable elements and development." *Cell* 56.2 (1989): 181-191.

Feschotte, Cédric. "Transposable elements and the evolution of regulatory networks." *Nature Reviews Genetics* 9.5 (2008): 397-405.

Fiston-Lavier, Anna-Sophie, et al. "T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data." *Nucleic acids research* 43.4 (2015): e22-e22.

Flutre, Timothée, et al. "Considering transposable element diversification in de novo annotation approaches." *PloS one* 6.1 (2011): e16526.

Freeling, Michael, et al. "Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants." *Current opinion in plant biology* 15.2 (2012): 131-139.

Fukai, Eigo, et al. "Establishment of a Lotus japonicus gene tagging population using the exon-targeting endogenous retrotransposon LORE1." *The Plant Journal* 69.4 (2012): 720-730.

Gao, Xiang, et al. "Chromodomains direct integration of retrotransposons to heterochromatin." *Genome research* 18.3 (2008): 359-369.

Gao, Dongying, et al. "A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes." *PloS one* 7.2 (2012): e32010.

Garcia-Mas, Jordi, et al. "The genome of melon (Cucumis melo L.)." *Proceedings of the National Academy of Sciences* 109.29 (2012): 11872-11877.

Gaut, Brandon S., et al. "Recombination: an underappreciated factor in the evolution of plant genomes." *Nature Reviews Genetics* 8.1 (2007): 77-84.

Gifford, Wesley D., Samuel L. Pfaff, and Todd S. Macfarlan. "Transposable elements as genetic regulatory substrates in early development." *Trends in cell biology* 23.5 (2013): 218-226.

Grandbastien, Marie-Angèle, Albert Spielmann, and Michel Caboche. "Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics." *Nature* 337.6205 (1989): 376-380.

Grandbastien, M-A., et al. "Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae." *Cytogenetic and genome research* 110.1-4 (2005): 229-241.

Guermonprez, Hélène, et al. "MITEs, miniature elements with a major role in plant genome evolution." *Plant Transposable Elements*. Springer Berlin Heidelberg, 2012. 113-124.

Han, Yujun, and Susan R. Wessler. "MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences." *Nucleic acids research* (2010): gkq862.

Hanin, Moez, and Jerzy Paszkowski. "Plant genome modification by homologous recombination." *Current opinion in plant biology* 6.2 (2003): 157-162.

Hénaff, Elizabeth, et al. "Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution." *BMC genomics* 16.1 (2015): 768.

Heyman, Jefri, et al. "Arabidopsis ULTRAVIOLET-B-INSENSITIVE4 maintains cell division activity by temporal inhibition of the anaphase-promoting complex/cyclosome." *The Plant Cell* 23.12 (2011): 4394-4410.

Hill, Alexander S., et al. "The most frequent constitutional translocation in humans, the t (11; 22)(q23; q11) is due to a highly specific Alu-mediated recombination." *Human molecular genetics* 9.10 (2000): 1525-1532.

Hirochika, Hirohiko. "Contribution of the Tos17 retrotransposon to rice functional genomics." *Current opinion in plant biology* 4.2 (2001): 118-122.

Hoen, Douglas R., et al. "A call for benchmarking transposable element annotation methods." *Mobile DNA* 6.1 (2015): 1.

Hormozdiari, Fereydoun, et al. "Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery." *Bioinformatics* 26.12 (2010): i350-i357.

Hou, Yi, et al. "Retrotransposon vectors for gene delivery in plants." *Mobile DNA* 1.1 (2010): 1.

Huang, Sanwen, et al. "The genome of the cucumber, Cucumis sativus L." *Nature genetics* 41.12 (2009): 1275-1281.

Ibarra-Laclette, Enrique, et al. "Architecture and evolution of a minute plant genome." *Nature* 498.7452 (2013): 94-98.

International Barley Genome Sequencing Consortium. "A physical, genetic and functional sequence assembly of the barley genome." *Nature* 491.7426 (2012): 711-716.

Ito, Hidetaka, and Tetsuji Kakutani. "Control of transposable elements in Arabidopsis thaliana." *Chromosome research* 22.2 (2014): 217-223.

Jacob, Yannick, et al. "ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing." *Nature structural & molecular biology* 16.7 (2009): 763-768.

Jacques, Pierre-Etienne, Justin Jeyakani, and Guillaume Bourque. "The majority of primate-specific regulatory sequences are derived from transposable elements." *PLoS Genet* 9.5 (2013): e1003504.

Ji, Jia-lei, et al. "Recessive male sterility in cabbage (Brassica oleracea var. capitata) caused by loss of function of BoCYP704B1 due to the insertion of a LTR-retrotransposon." *Theoretical and Applied Genetics* (2017): 1-11.

Jiang, Chuan, et al. "ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data." *BMC bioinformatics* 16.1 (2015): 72.

Jin, Weiwei, et al. "Maize centromeres: organization and functional adaptation in the genetic background of oat." *The Plant Cell* 16.3 (2004): 571-581.

Jouffroy, Ophélie, et al. "Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening." *BMC genomics* 17.1 (2016): 624.

Kaeppler, Shawn M., Heidi F. Kaeppler, and Yong Rhee. "Epigenetic aspects of somaclonal variation in plants." *Plant molecular biology* 43.2-3 (2000): 179-188.

Kalendar, Ruslan, et al. "Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes." *Genetics* 166.3 (2004): 1437-1450.

Kapitonov, Vladimir V., and Jerzy Jurka. "A universal classification of eukaryotic transposable elements implemented in Repbase." *Nature Reviews Genetics* 9.5 (2008): 411-412.

Kasajima, Ichiro, et al. "A protocol for rapid DNA extraction fromArabidopsis thaliana for PCR analysis." *Plant Molecular Biology Reporter* 22.1 (2004): 49-52.

Kazazian, Haig H. "Mobile elements: drivers of genome evolution." *Science* 303.5664 (2004): 1626-1632.

Ke, Ning, and Daniel F. Voytas. "cDNA of the yeast retrotransposon Ty5 preferentially recombines with substrates in silent chromatin." *Molecular and cellular biology* 19.1 (1999): 484-494.

Keane, Thomas M., Kim Wong, and David J. Adams. "RetroSeq: transposable element discovery from next-generation sequencing data." *Bioinformatics* 29.3 (2013): 389-390.

Kenrick, Paul, and Peter R. Crane. "The origin and early evolution of plants on land." *Nature* 389.6646 (1997): 33-39.

Monro, A. "Biosystematic Monograph of the Genus Cucumis (Cucurbitaceae), Botanical Identification of Cucumbers and Melons." JH Kirkbride Jr.; US Department of Agriculture. Boone: Parkway Publishers. 1993. 159pp. *Edinburgh Journal of Botany* 55.02 (1998): 325-326.

Kobayashi, Shozo, Nami Goto-Yamamoto, and Hirohiko Hirochika. "Retrotransposon-induced mutations in grape skin color." *Science* 304.5673 (2004): 982-982.

Krysan, Patrick J., Jeffery C. Young, and Michael R. Sussman. "T-DNA as an insertional mutagen in Arabidopsis." *The Plant Cell* 11.12 (1999): 2283-2290.

Kumar, Amar, and Jeffrey L. Bennetzen. "Plant retrotransposons." *Annual review of genetics* 33.1 (1999): 479-532.

Kumar, Amar, and Hirohiko Hirochika. "Applications of retrotransposons as genetic tools in plant biology." *Trends in plant science* 6.3 (2001): 127-134.

Kumar, Panqanamala Ramana, and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. Vol. 75. SIAM, 2015.

Kunarso, Galih, et al. "Transposable elements have rewired the core regulatory network of human embryonic stem cells." *Nature genetics* 42.7 (2010): 631-634.

Lamb, Jonathan C., et al. "Plant chromosomes from end to end: telomeres, heterochromatin and centromeres." *Current opinion in plant biology* 10.2 (2007): 116-122.

Lammens, Tim, et al. "Atypical E2Fs: new players in the E2F transcription factor family." *Trends in cell biology* 19.3 (2009): 111-118.

Le, Quang Hien, et al. "Distribution dynamics of the Tnt1 retrotransposon in tobacco." *Molecular Genetics and Genomics* 278.6 (2007): 639-651.

Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* 25.14 (2009): 1754-1760.

Li, Dawei, et al. "Syntenic relationships between cucumber (Cucumis sativus L.) and melon (C. melo L.) chromosomes as revealed by comparative genetic mapping." *BMC genomics* 12.1 (2011a): 396.

Li, Zhen, et al. "RNA-Seq improves annotation of protein-coding genes in the cucumber genome." *BMC genomics* 12.1 (2011b): 540.

Li, Jia-Yang, Jun Wang, and Robert S. Zeigler. "The 3,000 rice genomes project: new opportunities and challenges for future rice research." *GigaScience* 3.1 (2014): 8.

Lin, Rongcheng, et al. "Transposase-derived transcription factors regulate light signaling in Arabidopsis." *Science* 318.5854 (2007): 1302-1305.

Lin, Tao, et al. "Genomic analyses provide insights into the history of tomato breeding." *Nature genetics* 46.11 (2014): 1220-1226.

Lisch, Damon. "How important are transposons for plant evolution?" *Nature Reviews Genetics* 14.1 (2013): 49-61.

Liu, Sanzhen, et al. "Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome." *PLoS Genet* 5.11 (2009): e1000733.

Lou, Qunfeng, et al. "Integration of high-resolution physical and genetic map reveals differential recombination frequency between chromosomes and the genome assembling quality in cucumber." *PloS one* 8.5 (2013): e62676.

Lucas, H., et al. "RNA-mediated transposition of the tobacco retrotransposon Tnt1 in Arabidopsis thaliana." *The EMBO journal* 14.10 (1995): 2364.

Luo, Chongyuan, et al. "Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production." *The Plant Journal* 73.1 (2013): 77-90.

Lynch, Vincent J., et al. "Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals." *Nature genetics* 43.11 (2011): 1154-1159.

Maere, Steven, Karel Heymans, and Martin Kuiper. "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." *Bioinformatics* 21.16 (2005): 3448-3449.

Makarevitch, Irina, et al. "Transposable elements contribute to activation of maize genes in response to abiotic stress." *PLoS Genet* 11.1 (2015): e1004915.

Martin, Antoine, et al. "A transposon-induced epigenetic change leads to sex determination in melon." *Nature* 461.7267 (2009): 1135-1138.

Maumus, Florian, and Hadi Quesneville. "Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter." *PLoS One* 9.4 (2014): e94101.

Mazier, Marianne, et al. "Successful gene tagging in lettuce using the Tnt1 retrotransposon from tobacco." *Plant physiology* 144.1 (2007): 18-31.

McClintock, Barbara “Cytogenetic studies of maize and Neurospora.” *Carnegie Inst Washington Year Book* (1947) 46:146–152.

McClintock, Barbara “Mutable loci in maize.” *Carnegie Inst Washington Year Book* (1948) 47:155–169.

McCue, Andrea D., et al. "Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA." *PLoS Genet* 8.2 (2012): e1002474.

Meissner, Rafael, et al. "A high throughput system for transposon tagging and promoter trapping in tomato." *The Plant Journal* 22.3 (2000): 265-274.

Miyao, Akio, et al. "Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome." *The Plant Cell* 15.8 (2003): 1771-1780.

Momose, Masaki, Yutaka Abe, and Yoshihiro Ozeki. "Miniature inverted-repeat transposable elements of Stowaway are active in potato." *Genetics* 186.1 (2010): 59-66.

Naito, Ken, et al. "Unexpected consequences of a sudden and massive transposon amplification on rice gene expression." *Nature* 461.7267 (2009): 1130-1134.

Naouar, Naïra, et al. "Quantitative RNA expression analysis with Affymetrix Tiling 1.0 R arrays identifies new E2F target genes." *The Plant Journal* 57.1 (2009): 184-194.

Natali, Lucia, et al. "The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads." *BMC genomics* 14.1 (2013): 686.

Neumann, Pavel, et al. "Plant centromeric retrotransposons: a structural and cytogenetic perspective." *Mobile DNA* 2.1 (2011): 1.

Nystedt, Björn, et al. "The Norway spruce genome sequence and conifer genome evolution." *Nature* 497.7451 (2013): 579-584.

Okamoto, Hiroyuki, and Hirohiko Hirochika. "Efficient insertion mutagenesis of Arabidopsis by tissue culture-induced activation of the tobacco retrotransposon Tto1." *The Plant Journal* 23.2 (2000): 291-304.

Oliver, Keith R., Jen A. McComb, and Wayne K. Greene. "Transposable elements: powerful contributors to angiosperm evolution and diversity." *Genome biology and evolution* 5.10 (2013): 1886-1901.

Olsen, Kenneth M., and Jonathan F. Wendel. "A bountiful harvest: genomic insights into crop domestication phenotypes." *Annual Review of Plant Biology* 64 (2013): 47-70.

Pérez-Hormaeche, Javier, et al. "Invasion of the Arabidopsis genome by the tobacco retrotransposon Tnt1 is controlled by reversible transcriptional gene silencing." *Plant physiology* 147.3 (2008): 1264-1278.

Permal, Emmanuelle, Timothée Flutre, and Hadi Quesneville. "Roadmap for annotating transposable elements in eukaryote genomes." *Mobile Genetic Elements: Protocols and Genomic Applications* (2012): 53-68.

Peterson-Burch, Brooke D., Dan Nettleton, and Daniel F. Voytas. "Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae." *Genome biology* 5.10 (2004): 1.

Piacentini, Lucia, et al. "Transposons, environmental changes, and heritable induced phenotypic variability." *Chromosoma* 123.4 (2014): 345-354.

Piegu, Benoit, et al. "Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice." *Genome Research* 16.10 (2006): 1262-1269.

Piffanelli, Pietro, et al. "Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library." *Plant molecular biology* 65.5 (2007): 587-601.

Phytozome: a tool for green plant comparative genomics. [http://www. phytozome.net/]

Plomion, Christophe, et al. "Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies." *Molecular ecology resources* 16.1 (2016): 254-265.

Ponce De Leon, Ines, et al. "Physcomitrella patens activates reinforcement of the cell wall, programmed cell death and accumulation of evolutionary conserved defence signals, such as salicylic acid and 12-oxo-phytodienoic acid, but not jasmonic acid, upon Botrytis cinerea infection." *Molecular plant pathology* 13.8 (2012): 960-974.

Price, Alkes L., Neil C. Jones, and Pavel A. Pevzner. "De novo identification of repeat families in large genomes." *Bioinformatics* 21. Suppl 1 (2005): i351-i358.

Quesneville, Hadi, et al. "Combined evidence annotation of transposable elements in genome sequences." *PLoS computational biology* 1.2 (2005): e22.

Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26.6 (2010): 841-842.

Ramachandran, Srinivasan, and Venkatesan Sundaresan. "Transposons as tools for functional genomics." *Plant Physiology and Biochemistry* 39.3 (2001): 243-252.

Ramirez-Parra, Elena, Corinne Fründt, and Crisanto Gutierrez. "A genome-wide identification of E2F-regulated genes in Arabidopsis." *The Plant Journal* 33.4 (2003): 801-811.

Ramirez-Parra, Elena, et al. "Role of an atypical E2F transcription factor in the control of Arabidopsis cell growth and differentiation." *The Plant Cell* 16.9 (2004): 2350-2363.

Ramirez-Parra, Elena, et al. "E2F–DP transcription factors." *Annual Plant Reviews: Cell Cycle Control and Plant Development* 32 (2007): 138-163.

Rebollo, Rita, Mark T. Romanish, and Dixie L. Mager. "Transposable elements: an abundant and natural source of regulatory sequences for host genes." *Annual review of genetics* 46 (2012): 21-42.

Rensing, Stefan A., et al. "The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants." *Science* 319.5859 (2008): 64-69.

Repp, Alexander, et al. "Phosphoinositide-specific phospholipase C is involved in cytokinin and gravity responses in the moss Physcomitrella patens." *The Plant Journal* 40.2 (2004): 250-259.

Rishishwar, Lavanya, Leonardo Mariño-Ramírez, and I. King Jordan. "Benchmarking computational tools for polymorphic transposable element detection." *Briefings in Bioinformatics* (2016): bbw072.

Sakakibara, Keiko, et al. "Involvement of auxin and a homeodomain-leucine zipper I gene in rhizoid development of the moss Physcomitrella patens." *Development* 130.20 (2003): 4835-4846.

Saladié, Montserrat, et al. "Comparative transcriptional profiling analysis of developing melon (Cucumis melo L.) fruit from climacteric and non-climacteric varieties." *BMC genomics* 16.1 (2015): 440.

Salvi, Silvio, et al. "Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize." *Proceedings of the National Academy of Sciences* 104.27 (2007): 11376-11381.

Sambrook, Joseph, Edward F. Fritsch, and Tom Maniatis. *Molecular cloning*. Vol. 2. New York: Cold Spring Harbor laboratory press, 1989.

Sanseverino, Walter, et al. "Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome." *Molecular biology and evolution* (2015): msv152.

Santiago, Néstor, et al. "Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana." *Molecular biology and evolution* 19.12 (2002): 2285-2293.

Schaefer, D., et al. "Stable transformation of the moss Physcomitrella patens." *Molecular and General Genetics MGG* 226.3 (1991): 418-424.

Schaefer, Hanno. "Cucumis (Cucurbitaceae) must include Cucumella, Dicoelospermum, Mukia, Myrmecosicyos, and Oreosyce: a recircumscription based on nuclear and plastid DNA data." *Blumea-Biodiversity, Evolution and Biogeography of Plants* 52.1 (2007): 165-177.

Schaefer, D. G., et al. "RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens." *DNA repair* 9.5 (2010): 526-533.

Schmidt, Dominic, et al. "Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages." *Cell* 148.1 (2012): 335-348.

Schneeberger, Korbinian, et al. "Reference-guided assembly of four diverse Arabidopsis thaliana genomes." *Proceedings of the National Academy of Sciences* 108.25 (2011): 10249-10254.

Schuermann, David, et al. "The dual nature of homologous recombination in plants." *TRENDS in Genetics* 21.3 (2005): 172-181.

Schwartz, Amy, et al. "Reconstructing hominid Y evolution: X-homologous block, created by X–Y transposition, was disrupted by Yp inversion through LINE—LINE recombination." *Human Molecular Genetics* 7.1 (1998): 1-11.

Schween, G., et al. "Unique tissue-specific cell cycle in Physcomitrella." *Plant Biology* 5.01 (2003): 50-58.

Sebastian, Patrizia, et al. "Cucumber (Cucumis sativus) and melon (C. melo) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia." *Proceedings of the National Academy of Sciences* 107.32 (2010): 14269-14273.

Shen, Jianqiang, et al. "Translational repression by a miniature inverted-repeat transposable element in the 3′ untranslated region." *Nature Communications* 8 (2017).

Smit, Arian FA, Robert Hubley, and Phil Green. "RepeatMasker Open-3.0." (1996): 1996.

Smit, A. F. A., R. Hubley, and P. Green. "RepeatModeler Open-1.0. 2008-2010." *Access date Dec* (2014).

Studer, Anthony, et al. "Identification of a functional transposon insertion in the maize domestication gene tb1." *Nature genetics* 43.11 (2011): 1160-1163.

Sundaresan, Venkatesan, et al. "Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements." *Genes & development* 9.14 (1995): 1797-1810.

Sundaresan, Venkatesan. "Horizontal spread of transposon mutagenesis: new uses for old elements." *Trends in Plant Science* 1.6 (1996): 184-190.

Tenaillon, Maud I., Jesse D. Hollister, and Brandon S. Gaut. "A triptych of the evolution of plant transposable elements." *Trends in plant science* 15.8 (2010): 471-478.

This, Patrice, et al. "Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1." *Theoretical and Applied Genetics* 114.4 (2007): 723-730.

Tomato Genome Consortium. "The tomato genome sequence provides insights into fleshy fruit evolution." *Nature* 485.7400 (2012): 635-641.

Trouiller, Benedicte, et al. "MSH2 is essential for the preservation of genome integrity and prevents homeologous recombination in the moss Physcomitrella patens." *Nucleic acids research* 34.1 (2006): 232-242.

Untergasser, Andreas, et al. "Primer3—new capabilities and interfaces." *Nucleic acids research* 40.15 (2012): e115-e115.

Urbański, Dorian Fabian, et al. "Genome-wide LORE1 retrotransposon mutagenesis and high-throughput insertion detection in Lotus japonicus." *The Plant Journal* 69.4 (2012): 731-741.

VanBuren, Robert, et al. "The genome of black raspberry (Rubus occidentalis)." *The Plant Journal* 87.6 (2016): 535-547.

Van Den Heuvel, Sander, and Nicholas J. Dyson. "Conserved functions of the pRB and E2F families." *Nature reviews Molecular cell biology* 9.9 (2008): 713-724.

Vandepoele, Klaas, et al. "Genome-wide identification of potential plant E2F target genes." *Plant physiology* 139.1 (2005): 316-328.

Vendramin, Elisa, et al. "A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach." *PLoS One* 9.3 (2014): e90574.

Vitte, Clémentine, et al. "The bright side of transposons in crop evolution." *Briefings in functional genomics* (2014): elu002.

Walbot, Virginia. "Saturation mutagenesis using maize transposons." *Current opinion in plant biology* 3.2 (2000): 103-107.

Wang, Ting, et al. "Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53." *Proceedings of the National Academy of Sciences* 104.47 (2007): 18613-18618.

Wang, Jianrong, et al. "A c-Myc regulatory subnetwork from human transposable element sequences." *Molecular BioSystems* 5.12 (2009): 1831-1839.

Waterhouse, Peter M., and Christopher A. Helliwell. "Exploring plant genomes by RNA-induced gene silencing." *Nature Reviews Genetics* 4.1 (2003): 29-38.

Waugh, R., et al. "Genetic distribution of Bare–1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP)." *Molecular and General Genetics MGG* 253.6 (1997): 687-694.

Wicker, Thomas, et al. "A unified classification system for eukaryotic transposable elements." *Nature Reviews Genetics* 8.12 (2007): 973-982.

Willing, Eva-Maria, et al. "Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation." *Nature Plants* 1.2 (2015).

Wisman, Ellen, et al. "Knock-out mutants from an En-1 mutagenized Arabidopsis thaliana population generate phenylpropanoid biosynthesis phenotypes." *Proceedings of the National Academy of Sciences* 95.21 (1998): 12432-12437.

Witte, Claus-Peter, et al. "Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes." *Proceedings of the National Academy of Sciences* 98.24 (2001): 13778-13783.

Wong, Lee H., and KH Andy Choo. "Evolutionary dynamics of transposable elements at the centromere." *TRENDS in Genetics* 20.12 (2004): 611-616.

Xu, Zhao, and Hao Wang. "LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons." *Nucleic acids research* 35. Suppl 2 (2007): W265-W268.

Yang, Luming, et al. "Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly." *The Plant Journal* 71.6 (2012): 895-906.

Yang, Luming, et al. "Next-generation sequencing, FISH mapping and synteny-based modeling reveal mechanisms of decreasing dysploidy in Cucumis." *The Plant Journal* 77.1 (2014): 16-30.

Yang, Qin, et al. "CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize." *Proceedings of the National Academy of Sciences* 110.42 (2013): 16969-16974.

Yao, Jia-Long, Yi-Hu Dong, and Bret AM Morris. "Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor." *Proceedings of the National Academy of Sciences* 98.3 (2001): 1306-1311.

Ye, Kai, et al. "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." *Bioinformatics* 25.21 (2009): 2865-2871.

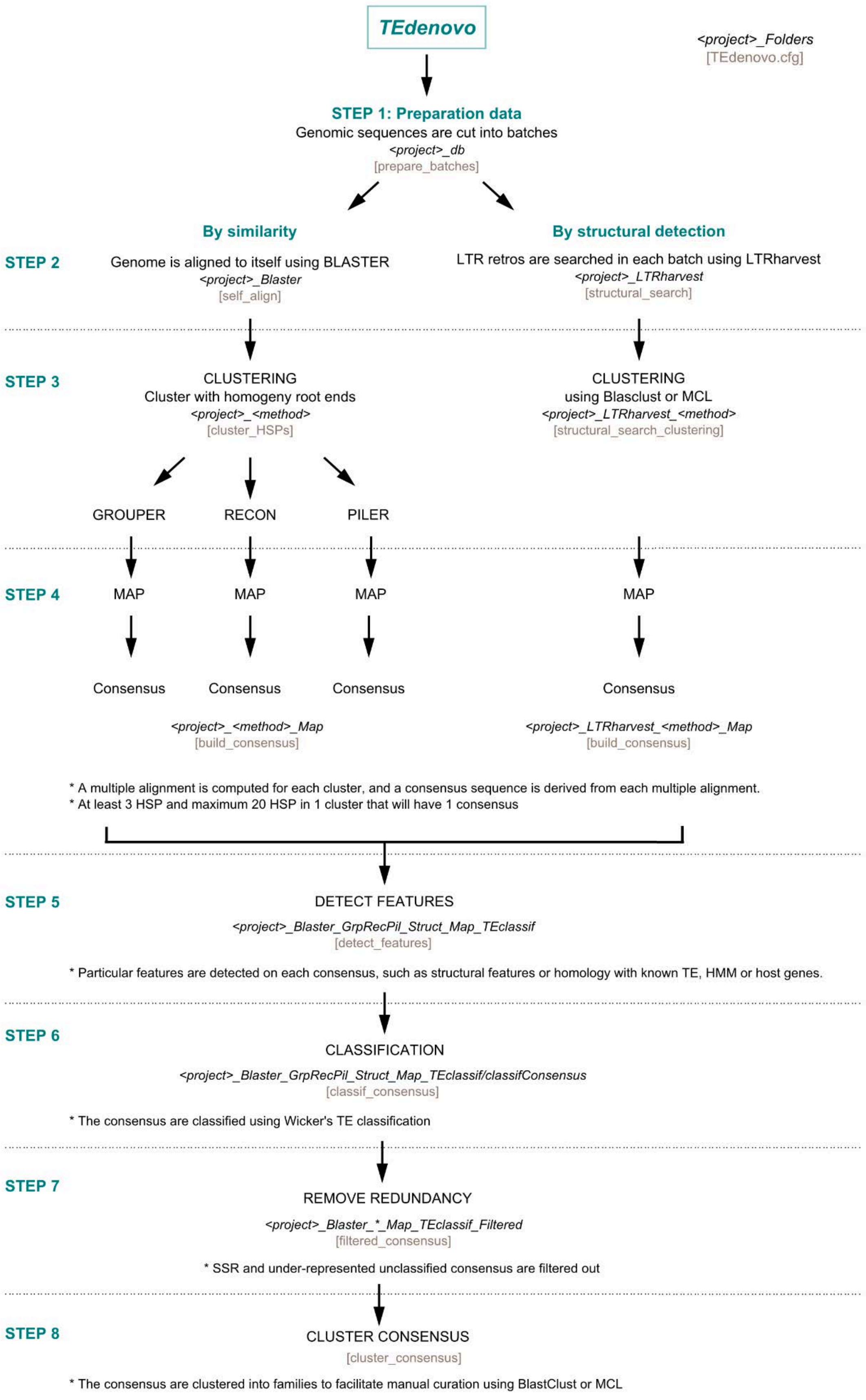
Zamudio, Natasha, et al. "DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination." *Genes & development* 29.12 (2015): 1256-1270.

Zhao, Meixia, and Jianxin Ma. "Co-evolution of plant LTR-retrotransposons and their host genomes." *Protein & cell* 4.7 (2013): 493-501.

Zhuang, Jiali, et al. "TEMP: a computational method for analyzing transposable element polymorphism in populations." *Nucleic acids research* 42.11 (2014): 6826-6838.

Annexes

Annex 1. *TEdenovo* pipeline of the REPET package, adapted from URGI (https://urgi.versailles.inra.fr/Tools/REPET)



* A multiple alignment is computed for each cluster, and a consensus sequence is derived from each multiple alignment.
* At least 3 HSP and maximum 20 HSP in 1 cluster that will have 1 consensus

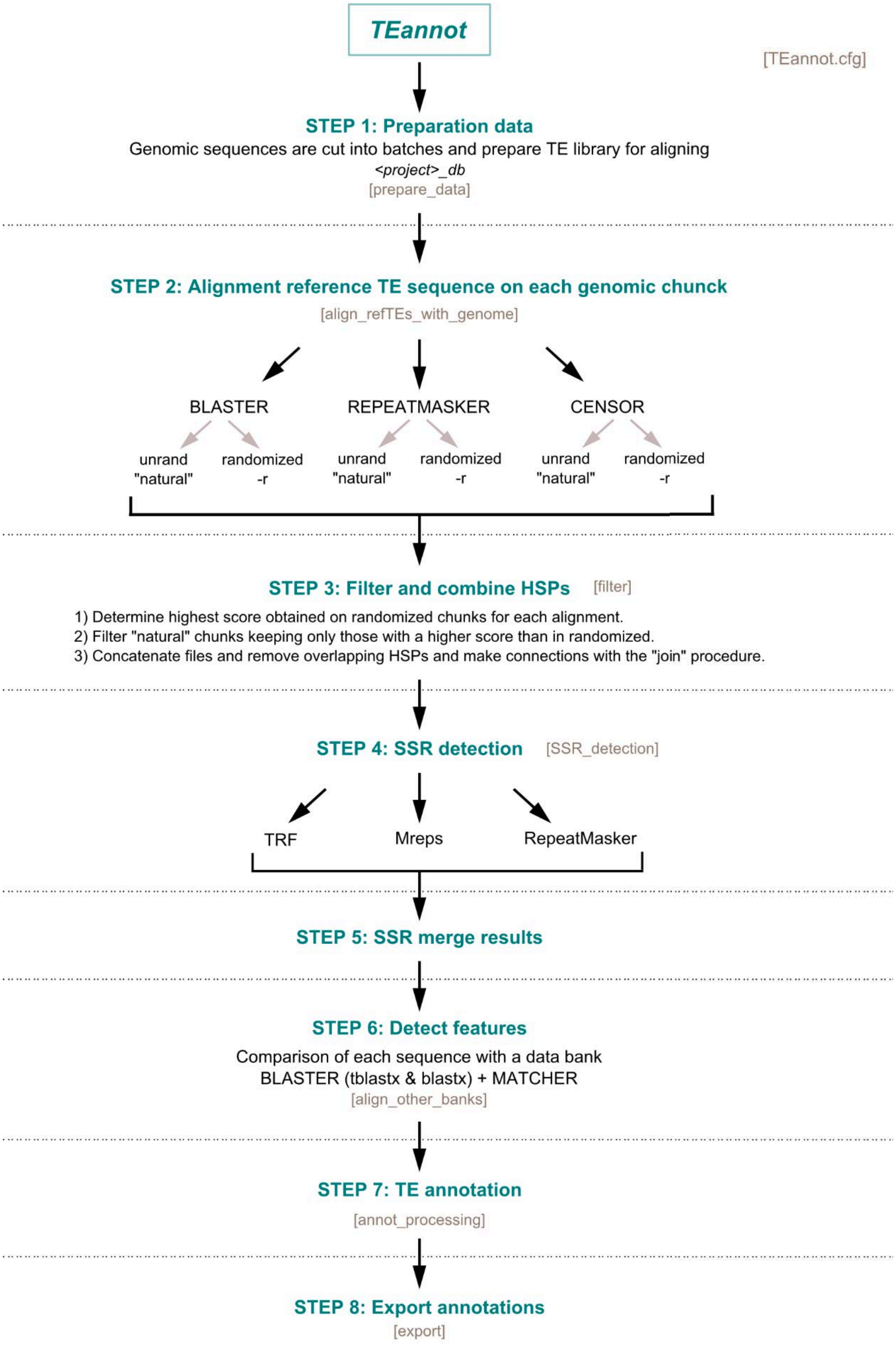
* Particular features are detected on each consensus, such as structural features or homology with known TE, HMM or host genes.

* The consensus are classified using Wicker's TE classification

* SSR and under-represented unclassified consensus are filtered out

* The consensus are clustered into families to facilitate manual curation using BlastClust or MCL

Annex 2. *TEannot* pipeline of the REPET package, adapted from URGI (https://urgi.versailles.inra.fr/Tools/REPET)



STEP 1: Preparation data

Genomic sequences are cut into batches and prepare TE library for aligning

STEP 2: Alignment reference TE sequence on each genomic chunck

STEP 3: Filter and combine HSPs

1) Determine highest score obtained on randomized chunks for each alignment.
2) Filter "natural" chunks keeping only those with a higher score than in randomized.
3) Concatenate files and remove overlapping HSPs and make connections with the "join" procedure.

STEP 4: SSR detection

STEP 5: SSR merge results

STEP 6: Detect features

Comparison of each sequence with a data bank
BLASTER (tblastx & blastx) + MATCHER

STEP 7: TE annotation

STEP 8: Export annotations

Annex 3. Scientific publications

Hénaff, Elizabeth, **Cristina Vives**, Bénédicte Desvoyes, Ankita Chaurasia, Jordi Payet, Crisanto Gutierrez, and Josep M Casacuberta. 2014. "Extensive Amplification of the E2F Transcription Factor Binding Sites by Transposons during Evolution of Brassica Species." The Plant Journal: For Cell and Molecular Biology 77 (6): 852–62. doi:10.1111/tpj.12434.

Contreras, Beatriz, **Cristina Vives**, Roger Castells, and Josep M. Casacuberta. 2015. "The Impact of Transposable Elements in the Evolution of Plant Genomes: From Selfish Elements to Key Players." In Evolutionary Biology: Biodiversification from Genotype to Phenotype, 93–105. Cham: Springer International Publishing. doi:10.1007/978-3-319-19932-0_6.

Sanseverino, Walter, Elizabeth Hénaff, **Cristina Vives**, Sara Pinosio, William Burgos-Paz, Michele Morgante, Sebastián E Ramos-Onsins, Jordi Garcia-Mas, and Josep Maria Casacuberta. 2015. "Transposon Insertion, Structural Variations and SNPs Contribute to the Evolution of the Melon Genome." Molecular Biology and Evolution. doi:10.1093/molbev/msv152.

Vives, Cristina, Florence Charlot, Corinne Mhiri, Beatriz Contreras, Julien Daniel, Aline Epert, Daniel F. Voytas, Marie-Angèle Grandbastien, Fabien Nogué, and Josep M. Casacuberta. 2016. "Highly Efficient Gene Tagging in the Bryophyte Physcomitrella Patens Using the Tobacco (Nicotiana Tabacum) Tnt1 Retrotransposon." New Phytologist. doi:10.1111/nph.14152.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) doi: 10.1111/tpj.12434

Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of *Brassica* species

Elizabeth Hénaff[1], Cristina Vives[1], Bénédicte Desvoyes[2], Ankita Chaurasia[1], Jordi Payet[3], Crisanto Gutierrez[2] and Josep M. Casacuberta[1,*]

[1]*Center for Research in Agricultural Genomics, Consejo Superior de Investigaciones Científicas-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona-Universitat de Barcelona, Campus Universitat Autònoma de Barcelona, Bellaterra – Cerdanyola del Vallès, 08193 Barcelona, Spain,*

[2]*Centro de Biologia Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas- Universidad Autónoma de Madrid, Cantoblanco, Nicolas Cabrera 1, 28049 Madrid, Spain, and*

[3]*Emili Badiella 49, 08225 Terrassa, Spain*

Received 31 October 2013; revised 24 December 2013; accepted 9 January 2014.

*For correspondence (e-mailjosep.casacuberta@cragenomica.es).

SUMMARY

Transposable elements (TEs) are major players in genome evolution. The effects of their movement vary from gene knockouts to more subtle effects such as changes in gene expression. It has recently been shown that TEs may contain transcription factor binding sites (TFBSs), and it has been proposed that they may rewire new genes into existing transcriptional networks. However, little is known about the dynamics of this process and its effect on transcription factor binding. Here we show that TEs have extensively amplified the number of sequences that match the E2F TFBS during *Brassica* speciation, and, as a result, as many as 85% of the sequences that fit the E2F TFBS consensus are within TEs in some *Brassica* species. We show that these sequences found within TEs bind E2Fa *in vivo*, which indicates a direct effect of these TEs on E2F-mediated gene regulation. Our results suggest that the TEs located close to genes may directly participate in gene promoters, whereas those located far from genes may have an indirect effect by diluting the effective amount of E2F protein able to bind to its cognate promoters. These results illustrate an extreme case of the effect of TEs in TFBS evolution, and suggest a singular way by which they affect host genes by modulating essential transcriptional networks.

Keywords: transposon, MITE, transcription factor binding site, evolution, *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Capsela rubella*, *Brassica rapa*, *Thelungella halopila*, transcriptional network.

INTRODUCTION

Transposable elements (TEs) are mobile genetic units that are present in all eukaryotes and account for the majority of their genome in most cases. Mutations generated by the movement of TEs are a major source of the variability required for evolution. Selection against deleterious mutations, as well as strategies evolved by TEs and their host genomes to minimize these unwanted consequences, ensure that most TE-induced mutations are neutral or only slightly deleterious. However, TEs are responsible for a large panoply of adaptive processes that have greatly contributed to genome evolution (Lisch, 2013). These include drastic changes such as production of new genes through the exaptation of TE-encoded gene functions, or more subtle effects such as modulation of expression of endogenous genes (Cowley and Oakey, 2013; Lisch, 2013). Indeed, insertions of TEs close to genes can modify their expression by providing new promoters, terminators or splice sites, as well as by attracting new combinations of epigenetic marks or providing target sequences for small regulatory RNAs (Cowley and Oakey, 2013; Lisch, 2013). Moreover, in the last few years, experimental evidence from ChIP analyses in animals, as well as computational approaches that aim to identify conserved regulatory sequences in animals, have shown that transcription factor binding sites (TFBSs) often co-localize with TEs. This is the case for master transcription factors, including p53, POU5F1, SOX2, c–Myc, CTCF, OCT4, NANOG and ERα (Wang *et al*., 2007, 2009; Bourque *et al*., 2008; Bourque, 2009; Kunarso *et al*., 2010; Lynch *et al*., 2011; Schmidt *et al*., 2012; Jacques *et al*., 2013). While not completely conclusive (de Souza *et al*., 2013), these experimental and computational data support the idea that insertion of TEs close to genes may allow extremely rapid modification of the regulation of sets of genes, thus rewiring transcriptional networks (Feschotte, 2008; Bourque, 2009).

E2F proteins are a family of transcription factors that play a key role in regulation of the cell cycle, DNA replication and development in both animals and plants (Ramirez-Parra *et al.*, 2007; van den Heuvel and Dyson, 2008; Lammens *et al.*, 2009). Eight E2F proteins have been described in mammals, and six (E2Fa–f) have been described in *Arabidopsis thaliana*. The DNA-binding domain(s) of all E2Fs are well conserved, and all E2F transcription factors bind sequences that fit the same consensus (DeGregori and Johnson, 2006). However, despite sharing the same binding site, E2F factors may be either transcriptional activators or repressors, and may have different target genes. In this way, they regulate various cell cycle-related processes such as apoptosis, DNA repair, cell proliferation, differentiation and development (Ramirez-Parra *et al.*, 2007; van den Heuvel and Dyson, 2008; Lammens *et al.*, 2009). Therefore, E2F factors form a complex network of transcriptional regulators with key and well-conserved functions in higher eukaryotes.

The results presented here show that a sequence that fits the consensus of the E2F binding site (BS) has been acquired by TEs that have amplified it to a high extent in various *Brassica* species. Our results show that these sequences bind an E2F factor *in vivo*, suggesting that they may participate in the E2F transcriptional network. We show that some transposons containing the E2F BS may participate in gene promoters, having rewired new genes into the E2F transcriptional network, while others may modulate E2F binding to its promoter sites by diluting the amount of E2F factor available.

RESULTS AND DISCUSSION

Transposons have amplified the E2F BS in various *Brassica* species

As part of the validation process of TE annotation tools using the *A. thaliana* genome as a test case, we serendipitously came across several miniature inverted-repeat transposable elements (MITEs) containing a sequence that was highly repeated in tandem. Analysis of this repeat revealed that it contains a sequence that fits the consensus for the E2F binding site (TTssCGssAA, where s = C or G; Ramirez-Parra *et al.*, 2003; Vandepoele *et al.*, 2005)). The number of E2F BSs in each element and the repeated nature of MITEs suggested that these E2F BSs embedded in TEs may account for a significant proportion of the E2F BSs present in the genome of *A. thaliana*. In order to investigate this, we identified all the sequences fitting the E2F BS consensus, and compared this annotation with the available annotation of TEs in *A. thaliana* (Ahmed *et al.*, 2011). This analysis showed that an unexpected 73% of the sequences fitting the consensus for the E2F BS are within an annotated TE. This is much higher than expected for a random distribution over the genome, as the annotated TEs account for 21% of the *A. thaliana* genome (Ahmed *et al.*, 2011). Moreover, we analyzed the distribution of other well-known plant TFBS, such as IBox, UP1, Gbox and MSA, using the same approach, and found that none are found in TEs at a frequency higher than expected for a random distribution, and that they are in fact under-represented in the TE fraction in most cases (Table 1).

Analysis of the occurrences of all sequences fitting the E2F BS consensus in the TE and non-TE fraction of the genome showed that only one of them, the sequence TTCCCGCCAA, is concentrated in TEs. This sequence is found in *A. thaliana* at a much higher number than the other sequences fitting the consensus (at least 14 times) (Figure 1). Moreover, 90% of the instances of this sequence are found within TEs (Figure 1 and Table 2). These results strongly suggest that TEs have amplified the sequence TTCCCGCCAA in *A. thaliana*. In order to determine whether this result is specific to this genome, we performed a similar analysis in four related *Brassica* species (*Arabidopsis lyrata*, *Capsela rubella*, *Brassica rapa* and *Thelungiella halophila*), as well as in the distantly related *Oryza sativa*. The results in Figure 2 and Table 2 show that the sequence TTCCCGCCAA, and only this sequence out of the various sequences fitting the consensus for the E2F binding site, is present at a much higher frequency per Mb than the other eight sequences in *A. thaliana*, *A. lyrata*, *C. rubella* and *B. rapa*. In contrast, all nine sequences are found at a similar frequency in rice (Figure 2 and Table 2). This sequence is not found at a significantly higher

Table 1 TFBS distribution in *A. thaliana* TEs, showing the number of instances of the sequences fitting the consensus of various TFBSs in *A. thaliana* over the whole genome or in the TE fraction of the genome

Box	Sequence	Number in genome	Number in TEs	Percentage in TEs
E2F	TTssCGssAA	2566	1874	73
Ibox	CTTATCCN	12 779	2672	20.9
UP1	GGCCCANN	22 693	3753	16.5
Gbox	GCCACGTN	6283	659	10.4
MSA	GACCGTTN	6031	988	16.4

number in *T. halophila*, and its frequency is similar to the frequencies observed in rice (Figure 2 and Table 2). In order to determine whether TEs are responsible for the higher frequency of the TTTCCCGCCAA sequence in all *Brassica* species, we analyzed the distribution of the various sequences fitting the E2F consensus among the TE and non-TE fraction of these genomes. TE annotations are available for *A. thaliana* (Ahmed *et al.*, 2011), *A. lyrata* (Hu *et al.*, 2011) and *O. sativa* (Ouyang *et al.*, 2007), but not for the other *Brassica* species. For this reason, we annotated them using RepeatMasker using the *A. thaliana* TE database, as was done for *A. lyrata* (Hu *et al.*, 2011). Interestingly, in all cases where the sequence TTCCCGCCAA is found at a higher copy number it is highly enriched in the TE fraction of the genome (Table 2), suggesting that the TEs have amplified it. *A. lyrata* could represent an extreme case as the sequence TTCCCGCCAA is found at least 75x more instances than any other sequence fitting the E2F consensus and is located in TEs in more than 94% of the instances (Table 2). This may be explained by the lower efficacy of TE silencing in *A. lyrata*, which has allowed TEs to proliferate to a higher extent in this genome (Hu *et al.*, 2011).

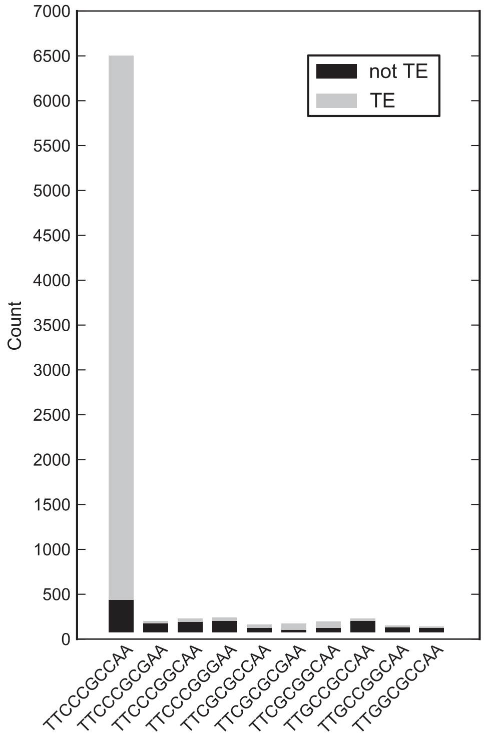


Figure 1. Number of instances of each of the ten sequences fitting the E2F consensus in the TE and non-TE fractions of the genome of *A. thaliana*.

The results presented here show that transposons have captured a sequence containing the binding site for the E2F transcription factor during evolution of some *Brassica* genomes. The effect of transposons on the evolution of gene regulation in eukaryotes has already been suggested (Medstrand *et al.*, 2005; Feschotte, 2008), and it is now clear that transposons contain and mobilize TFBSs (Wang *et al.*, 2007, 2009; Bourque *et al.*, 2008; Bourque, 2009; Kunarso *et al.*, 2010; Lynch *et al.*, 2011; Schmidt *et al.*, 2012; Jacques *et al.*, 2013). Here we report not only the capture but also extreme amplification of a TFBS during a relatively short evolutionary time.

Various TE families have amplified the E2F BS

In order to obtain insight into how this sequence was acquired, we sought to determine which TEs were responsible for the amplification observed. In each genome, a reduced number of families contain the majority of the E2F sequences (more than 80% of the E2F BSs in TEs in a given genome). Figure 3 shows the relative contribution of these main families for the five *Brassica* genomes.

The results show that six TE families are responsible for amplification of the E2F BS across the five *Brassica* species. Some families, such as *Simpleguy*1, have amplified the E2F BS in all genomes, while others, such as *Al*1,

Table 2 Distribution of sequences fitting the E2F BS consensus in TEs, showing the number of instances per Mb and the percentage found in the TE fraction of the genome (shown in parentheses) for each of the 10 sequences fitting the E2F consensus in six plant genomes

	Number of instances per Mb (% in TEs)					
	A. thaliana	*A. lyrata*	*C. rubella*	*S. rapa*	*T. halophila*	*O. sativa*
TTCCCGCCAA	15.57 (89.82)	53.95 (94.32)	19.76 (91.12)	24.58 (88.05)	3.26 (33.16)	2.54 (4.62)
TTCCCGCGAA	0.75 (29.21)	1.03 (19.51)	0.56 (14.93)	1.90 (39.38)	1.06 (25.40)	2.20 (17.56)
TTCCCGGCAA	0.99 (16.10)	1.32 (23.57)	0.76 (13.19)	2.05 (18.03)	1.96 (24.46)	2.02 (3.32)
TTCCCGGGAA	0.97 (14.66)	1.39 (21.08)	1.21 (15.97)	1.64 (15.90)	2.26 (32.34)	1.34 (8.13)
TTCGCGCCAA	0.62 (40.54)	0.75 (39.33)	0.67 (32.50)	0.98 (19.66)	0.93 (36.04)	3.20 (14.96)
TTCGCGCGAA	0.07 (25.00)	0.81 (72.16)	0.24 (0.00)	0.28 (33.33)	0.18 (9.09)	0.50 (1.69)
TTCGCGGCAA	0.44 (7.69)	1.02 (57.02)	0.46 (14.55)	0.73 (12.64)	0.90 (31.78)	1.91 (13.22)
TTGCCGCCAA	1.08 (34.88)	1.27 (15.89)	1.11 (24.24)	1.89 (14.22)	1.85 (11.82)	3.88 (4.11)
TTGCCGGCAA	0.28 (15.15)	0.61 (26.03)	0.33 (2.56)	0.79 (7.45)	0.55 (15.38)	2.10 (4.80)
TTGGCGGCAA	0.77 (60.87)	0.53 (19.05)	0.32 (18.42)	0.86 (12.62)	0.72 (15.12)	4.41 (14.86)

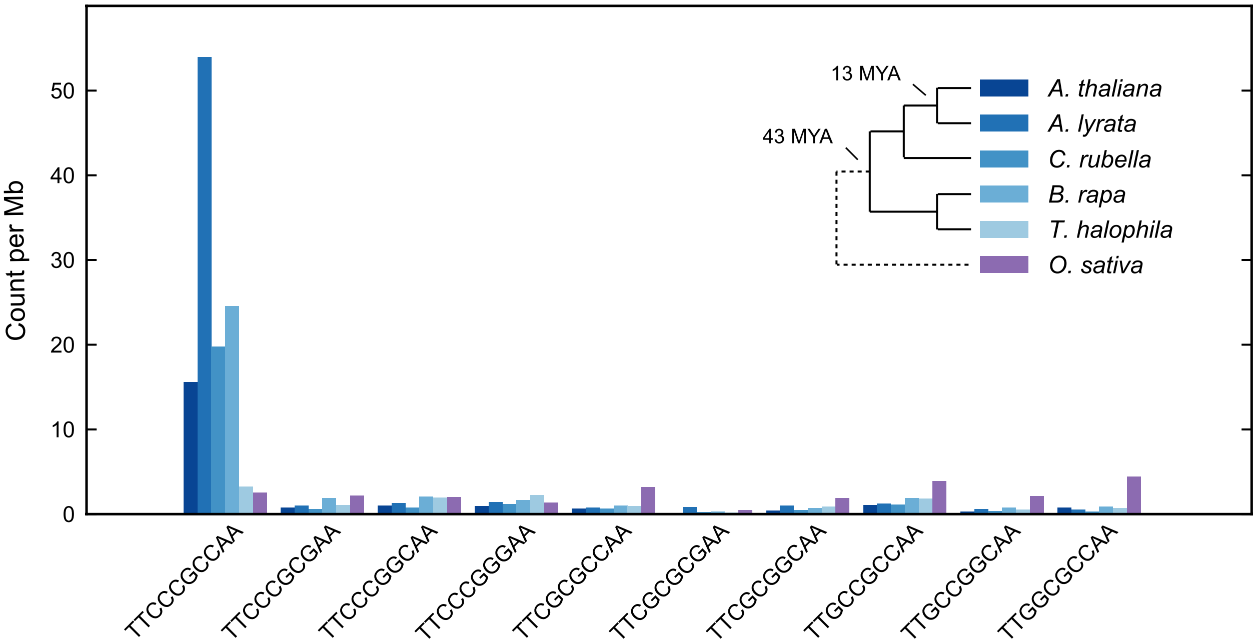


Figure 2. Number of instances per Mb of each of the ten sequences fitting the E2F consensus in six plant genomes.
Dendogram showing the phylogenetic relationships of the species analyzed, with an estimation of the divergence time of some nodes according to Beilstein *et al.* (2010). The phylogenetic relationship of *O. sativa*, used as an outgroup, to the other genomes analyzed is indicated by a dashed line.

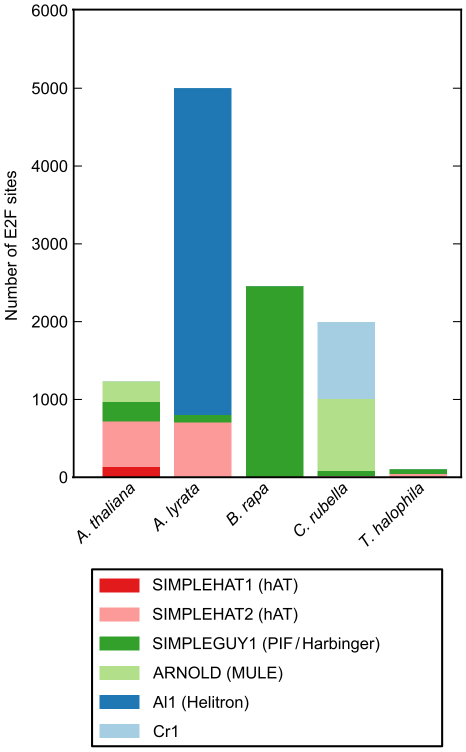


Figure 3. Number of E2F sites contributed by each TE family in the five *Brassica* genomes analyzed (only the major TE families collectively accounting for at least 80% of the E2F sites found in TEs in each genome are shown).
The super-family to which each TE family belongs is shown in parentheses, except for the *Cr*1 family whose complex nature precluded its classification.

seem to have been particularly active in just one of them. However, although they do not always represent a significant proportion of the TEs containing the E2F BS, all six families are present in all five genomes, suggesting that they were already present in their common ancestor and have been amplified to various degrees after the lineages split. It is interesting to note that *T. halophila* does contain TEs with the E2F BS, but these appear not to have been amplified, in contrast to the other four *Brassica* genomes in which TEs have greatly increased the number of E2F BSs.

The six TE families that have amplified the E2F BS in *Brassica* are related to at least four super-families of class II transposons (hAT, PIF/Harbinger, MULE and Helitron), and form sequence-uniform clusters of short non-coding elements that can be considered as MITEs. As already mentioned, in recent years, evidence has accumulated supporting an effect of TEs on mobilizing TFBSs and rewiring transcriptional networks. In most cases, the regulatory elements required for TE expression are co-opted by endogenous genes after insertion of the TE in gene-proximal regions (Cowley and Oakey, 2013). For this reason, the most frequent elements contributing to endogenous promoters are those that must be transcribed in order to transpose (i.e. retrotransposons) and that therefore contain a functional promoter that may in some case be co-opted as an alternative promoter (Testori *et al.*, 2012). MITEs are defective class II elements that transpose by a cut-and-paste mechanism that does not involve

SIMPLEHAT2 (hAT)

```
TAGGGGTGTCAAAATAGCTCAAAATTCATGGATCAACTCAACTCAACTCA
ACCCATAACCCTAATGAGTTGAAAATTTTGACTCCAATGAGTTTATGGGT
CAAATGAGTTATTAAGTCAATTGGTTTAATGAGTAAAATGAGTTGGGTTA
TAATGGTTAATGGTTTACCCAATTAACCCATCAAGTTTTATAAATTGAAT
TAAACCAACTAAAATCTTTAAACCAATGTCAATCTAAGTTTAACCAACAC
ATCTAAACCAATTTAATAAAATCATTTTTTTCCAAATTTCTTAAATATAC
AAGCGATGAAATTGAGAAAAAGTATACTCG

TAAT TTTTCCA CCAAAAAA CATAAACCCG
TAA  TTTTCCCGCCAAAAA  CGTAAACCCA
TGAT TTTTCC GCCAAAAA  CGTAAACCCG
TGAT A TTCCCGCCAAAAA  CGTAAACCCG
TAAT TTTTCC GCCAAAAA  CGTAAACCCG
TAA  TTTTCCCGC AAAAAA CATAAACCCA
TGAT TTTTCC GCCAAAA   CGTAAATCCG
TAA  TTTTCCCACCAAAAAA CGTAAACCCG
TGA  TTTTCTCGC AGAAA  CGTAAACACG
TAA  TTTTCCGGCCAAAAAAACGTAAACCCG
TGA  TTTTCTCGCCAAAAA  CGTAAATC
TGTAATTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAAA  CATAAACCCA
TGA  TTTTCCCGCCAAAA   CGTAAACC
TGTAATTTTCCCGCCAAAAAA  GTAAACCCG
TGA  TTTTTCTGC AGAAA  CGTAAACCCG
TAA  TTTTTCCGC AAAAAAACGTAAACCCG
TGA  TTTTCCCGCCAAAA   CATAAACC
TGTAATTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAAA  CGTAAGCCCGG
TAA  TTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTTCTGC AGAAA  CGTAAACCCG
TAA  TTTTTCCGC AAAAAAACGTAAACCCG
TGA  TTTTCCCGCCAAAA   CGTAAACCTG
TAA  TTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAA   CGTAAACCTG
TAA  TTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAAA  CGTAAGCCCGG
TAA  TTTTCCCGCCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAAA  CATAAACCTG
TAA  TTTTCCCGTCAAAAA  CGTAAACCCG
TGA  TTTTCCCGCCAAAAA  CGTAAACCCA
TAA  TTTTCCCGCCGAAAA  CATAAACC

TATAAAAATTAGAATCCGTAAATATTCTAAGTTTGATGATAATGAATTAA
TAATTATTAATAATTATAATTATTTATTATTGTTTTATAATAATTATTAA
TTAAATTATTACGTAAATTGGTTAAACCGTTTAACAACTCAACCCATCAA
ATTAAATGAGTTATGGGTTGACGCAACTCATTTTATTAAATGGGTTGGGT
CAACCTATAACTCATTTAATCATAAACTCATTTGATTATGAGTTAAGTTG
AGTTGGGCTACCTATTTTGACACCCCTA
```

SIMPLEGUY1 (PIF/harbinger)

```
GGCCATGTTTGTTTGTCCATCCACATGATCCATCCAAATGGAGATGAAAA
ATGATGTTTGTTTTTGACAATTTAACAAGCATTTGGATGGATCATCTGGA
TGATACATCTGAATGAGTTTCTAAATTTAGGCAAATTTATGAAACTCATT
TTGATGAATTTGATTAAACCATTTGGATGGAGATGGTTCATCCAAAAATA
ACAAAAGACTGATATACCCTCATTTATAGTCATAATTTAAATAATTTTAT
TTTCTATAGATTTCTTATTTAATATTTTTTAAATTACATATTGAAACAAA
AAAAACTCAAAAAATCATGAAATCACGGATTATTCTTTTATGTCAAAATC
GTAAAATTGTGGAATTGTGTTTTTTGGCCAAAAACCGCAAAATCATAATT
TTTGATGTAAAATTATAGAAATCGCTATTCCACCTAAAATCATAAAACTA
TATTTACGCTGAAACCACTAAATTATATTTTCTTGCCAAAATGACGATAT
TA

TG TTTTCCCGCCAAAACTGCAAAA TCA
TG TTTTCCTGCCAAAATTGCAAAA TCA
T  TTTTCCCGCAAAA          TCA
TG TTTTCCCGCCAAAACCACAAAAATCT
TA TTTTTACGCCAAAAACGCAAAA TCA
TGTTTTTCCCGCCAAAACCATAAAA TCA
TG TTTT CCGCCAAAACCGTAAAA TCA
TG TTTTTTCGC AACACCGTAAAA TTA
TG  TTTCCCGCCAAAACCGCAAAA TCA
TGTTTTTCCCGCCCAAACCGTAAAA TCA
TGTTTTTCCCGCCAAAACCGCAAAA TCA
TGTTTTTCCCGCCAAACCGCAAAA  TCA
TGTTTTCCTCGCCAAAACAGTAAAA TCA
TG TTTTCCCGCCAAAACCATAAAA TCA
TG TTTT  CGC AAATCCGTAAAA TCA
TG TTTTCCCGCCAAAACCGCAAAC TCA
TG TTTTCCCGCCAAAATCGCAAAA TCG
TA TTTTACTGCCAAAACCGTAAAA TCG
TGTATTCTCAG CAAA CTGTAAAA TCA
TA TTTTTCCGCCAAAACCGTAAAA TCG
TA TTTTCCCGCCAAAACCGTAAAA TCA
TA TTTTCCCGCCAA

TATGTTTGGTATTGATAATGGGTAATAATGTTGAATATTTGTCATTTGTT
AAAATGAAAACATGGGTATTTTAGTCATTTTTATCTAAATGAAAATGTGA
ATGAACTACAAAATGAAAATGTGAATGAACTACAAAATGATCAAACAAAT
AAAGATCCATTTAAATGCATCATTTAGATTAAATTACAAATAATGAAACA
AACATCATTCAAATGTATCATTCAAATGAATGTAAGTGGATCATTTGAAT
GGAAATGATAAATGATGAAACAAACAGGGCC
```

Figure 4. Examples of the sequence of two E2F-TEs belonging to the *Simplehat*2 and *Simpleguy*1 families.
The sequences of the terminal inverted repeats are shown in blue. The E2F BS sequences (TTCCCGCCAA) are shown in red. The central part of the sequences is aligned to show its tandemly repeated mini-satellite structure.

transcription of the MITE itself (Guermonprez *et al.*, 2012). Therefore, in contrast to most cases reported so far, the TFBSs contained in the MITEs reported here may have an effect on transcriptional regulation of endogenous genes without having a previous role in expression of the TE itself.

Closer inspection of these TEs shows that the E2F BS is always found repeated in the central part of the TE, and is frequently part of a longer motif that is repeated in tandem (up to 35 times), forming a mini-satellite (Figure 4). In order to determine whether the E2F motifs in TEs are generally found in a mini-satellite context in *A. thaliana*, we identified tandem repeats genome-wide using TRF (Benson, 1999). Figure 5 shows the whole set of E2F sites in the *A. thaliana* genome, and the percentage found within a TE, a mini-satellite, both or neither. These results show that, whereas most E2F BSs found within TEs in *A. thaliana* are associated with a mini-satellite structure, those found outside TEs are usually isolated. The double association of the E2F BS with MITEs and mini-satellites may explain their extensive amplification in *Brassica*. Indeed, on one hand, mini-satellites can easily expand, increasing the number of tandemly repeated motifs (Richard *et al.*, 2008), and, on the other hand, MITEs are

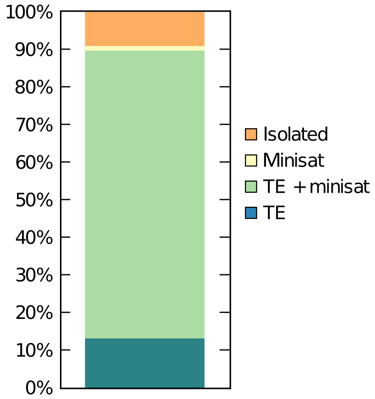


Figure 5. Percentage of total E2F sequences in the *A. thaliana* genome found within a TE, a mini-satellite, both or neither.

present as large families of elements (Guermonprez *et al.*, 2012). Most MITEs are thought to be the result of amplification of deletion derivatives of DNA transposons by an unknown amplification mechanism (Guermonprez *et al.*, 2012). According to this scenario, acquisition of an E2F BS by a single DNA transposon may give rise to an entire MITE family containing the E2F BS, which, if present in a mini-satellite structure, may also increase its copy number within each MITE. Interestingly, for at least one of the MITE families containing E2F BSs (*Simplehat*2 family), we have detected a longer copy (At4G05510) that may be the precursor of the entire family. This copy, which we have named hAT2, has the coding potential of a hAT transposase, and has high sequence similarity with *Simplehat*2 MITEs in the first 274 and last 357 of its 4045 nucleotides. Interestingly hAT2 contains three copies of the *Simplehat*2 mini-satellite, one of which contains the sequence TTCCCGCCAA. This suggests that the hAT2 transposon captured the mini-satellite and the E2F BS embedded in it, and that a deletion derivative of hAT2 amplified to generate the *Simplehat*2 family, which comprises 79 elements in *A. thaliana* that include 583 copies of the TTCCCGCCAA sequence (Figure S1).

A proportion of E2F-TEs are found in upstream proximal regions of genes

The most obvious effect that a TE containing a TFBS may have is to affect gene expression by insertion into a gene promoter. For this reason, we investigated the position of the E2F BS within or outside TEs with respect to genes in Arabidopsis. As a basis of comparison, we also assessed the distance to genes of four TFBSs that are not associated with TEs (see Table 1). The four TFBSs that are not related to TEs show a very similar distribution, with the vast majority of them found very close to the 5′ region of *A. thaliana* genes (Figure 6). This is consistent with the fact that plant promoters, particularly those of *A. thaliana*, which is a compact genome (Kaul *et al.*, 2000), do not usually contain distantly located regulatory elements. The distribution of the E2F BSs that are not associated with TEs does not differ from that of the TFBSs chosen as controls, with a clear concentration in the proximal upstream regions of genes (Figure 6), confirming previous reports (Ramirez-Parra *et al.*, 2003). However, only a small proportion of the E2F BSs located within TEs are found in these regions. The majority of them are located far from genes and are more evenly distributed between the upstream and downstream regions (Figure 6). This analysis strongly suggests that only a proportion of the E2F-TEs participate in gene promoters.

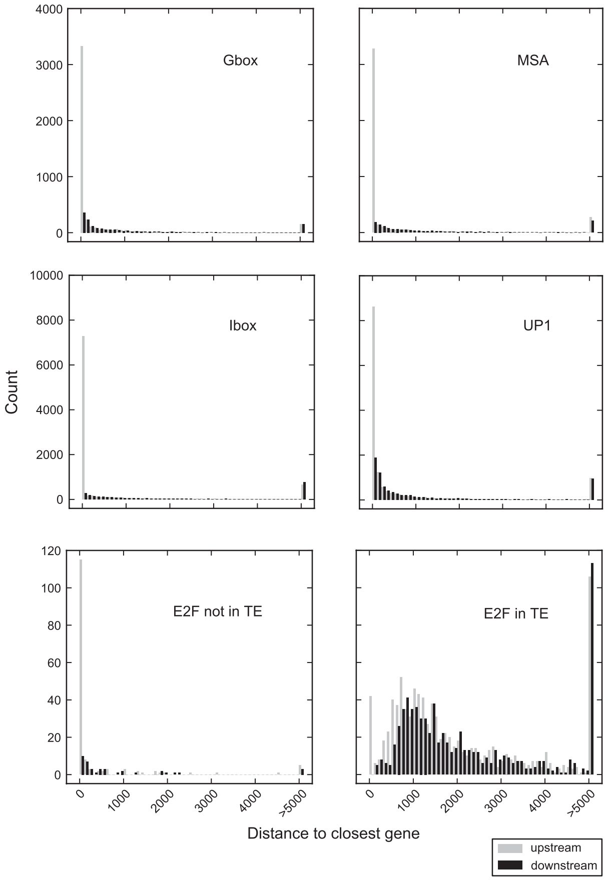


Figure 6. Number of instances of various TFBSs with respect to the distance to genes in *A. thaliana*.
Separate analyses of E2F BSs found within or outside of TEs are shown.

The members of the E2F family of transcription factors have various activation/repression roles on cell cycle-related target genes (Ramirez-Parra *et al.*, 2007; Lammens *et al.*, 2009) but share a DNA-binding domain and therefore the binding site (Zheng *et al.*, 1999). Insertion of E2F-TEs close to genes may potentially place those genes under the regulation of any E2F transcription factors. Previous studies have identified potential E2F-regulated genes by analyzing changes in expression in plants over-expressing both the E2Fa protein and its co-regulator DPa in two separate experiments using two types of microarrays (Vandepoele *et al*., 2005; Naouar *et al*., 2009). The combined data identified 1141 potential target genes that are over-expressed in the E2Fa-DPa over-expressing transgenic line (Naouar *et al*., 2009). Analysis of the 1 kb upstream region of those genes showed that 542 of them have an E2F BS in their potential promoter sequences (Takeda *et al*., 1999), suggesting that they may be E2F primary targets, and confirming previous observations (Ramirez-Parra *et al*., 2003). We searched these regions for E2F BSs and E2F-TEs, and found that, in five cases, the only E2F BSs within the first 1000 nucleotides are those contributed by an E2F-TE. These are cases in which an E2F-TE may contribute a regulatory sequence to a nearby gene, illustrating their potential for rewiring new genes into the E2F transcription network.

E2F binds E2F-TEs *in vivo*

In order to obtain further insight into the capacity of E2F-TEs to modulate E2F transcriptional regulation, we analyzed the chromatin with which these elements are associated, as well as their capacity to bind the E2F transcription factor *in vivo*. To this end, we used whole-genome ChIP-seq data for the histone modifications H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K9Ac and H3K9me2 (Luo *et al*., 2013). We wished to determine whether there is any particular mark that is significantly correlated with E2F sites either inside or outside TEs. We performed a x^2 test to evaluate the correlation of the epigenetic marks, and found that H3K27me1 is significantly correlated with E2F sites in TEs ($P = 1.8e\text{-}18$) and H3K4me2 is significantly correlated with E2F sites outside TEs ($P = 4.5e\text{-}60$). The other marks did not show significant correlation with either type of E2F site. As H3K27me1 is associated with heterochromatin and silencing in Arabidopsis (Jacob *et al*., 2009), this results suggests that E2F-TEs are associated with local heterochromatin. In order to confirm these data, we performed ChIP analyses using commercial antibodies raised against these histone modifications. The repetitive nature of the E2F-TEs makes it necessary to design one PCR primer in the non-repetitive region flanking the element to ensure specificity, while the second primer can be designed within the element itself, such that the amplified fragment contains the E2F BS. These requirements, and the size of the E2F-TEs, result in PCR fragments containing E2F TFBSs that are far too long for a quantitative PCR approach, and we therefore analyzed the ChIPs by semi-quantitative PCR. These analyses confirmed that isolated E2F BSs found in promoters of known E2F regulated genes are associated with a high level of H3K4me2 and a low level of H3K27me1, whereas most E2F-TEs have the opposite combination of epigenetic marks, with a high level of H3K27me1 and a low level of H3K4me2 (Figure 7 and Figure S2). Nevertheless, a small number of E2F-TEs inserted close to or within genes showed the same chromatin structure as the neighboring genes, characterized by a high level of H3K4me2 and a low level of H3K27me1 (Figure 7 and Figure S2). These results show that the chromatin that the E2F BS is associated with depends on whether the E2F BS is isolated or contained in a TE, as well as on the position of the E2F-TE with respect to genes. In general, isolated E2F BSs are associated with euchromatic epigenetic marks whereas E2F-TEs are associated with heterochromatic marks, with the exception of some E2F-TEs inserted within or close to genes. This result suggests that binding of E2F to various types of E2F BSs may be differentially regulated.

In order to obtain a direct insight into the binding of E2F *in vivo*, we performed ChIP analyses using an antibody raised against the E2Fa transcription factor (Heyman *et al*., 2011). We analyzed a total of 26 E2F-TEs (Tables S1 and S2) and four known E2F target genes as positive controls in at least two independent ChIP assays. We also analyzed the binding to two negative control sequences, a TE belonging to the *Simplehat*2 family of E2F-TEs whose

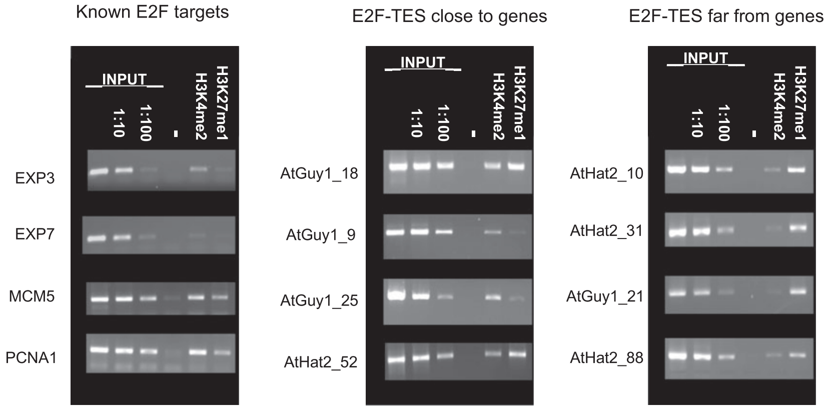


Figure 7. ChIP analyses of the epigenetic marks H3K4me2 and H3K27me1 associated with various types of E2F BS.

Four examples of each of the three classes of E2F BS [known E2F target genes (left), E2F-TEs close to genes (middle) and E2F-TEs far from genes (right)] are shown. A negative control (−) was performed immunoprecipitating with no antibody.

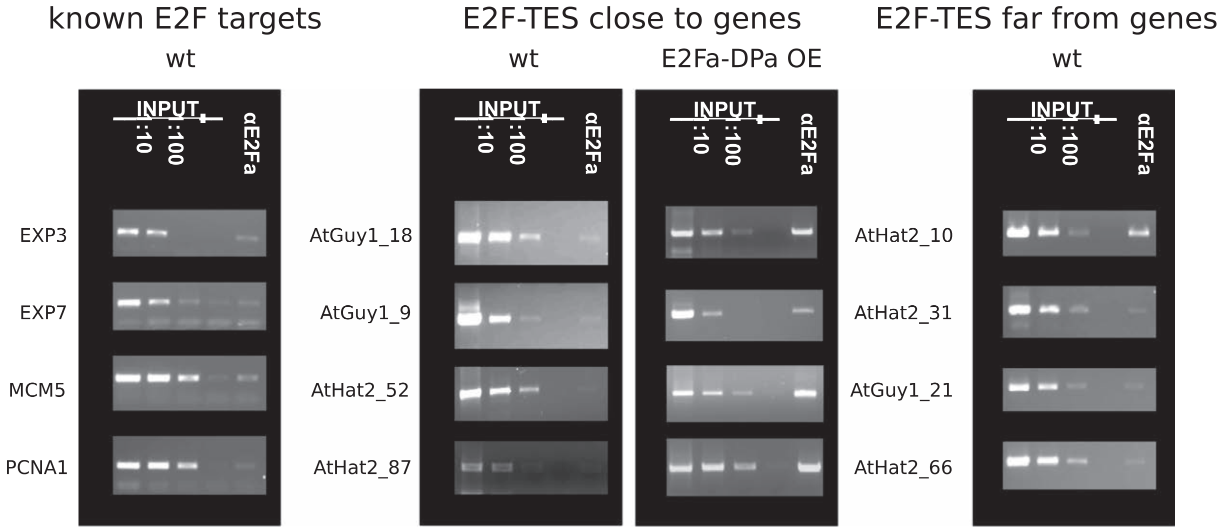


Figure 8. ChIP analyses of binding of the E2F protein to E2F BSs.
Four examples of each of the three classes of E2F BS [known E2F target genes (left), E2F-TEs close to genes (middle) and E2F-TEs far from genes (right)] in wild-type plants (wt) or plants over-expressing E2Fa and DPa transcription factors (E2Fa-DPa OE) are shown. A negative control (−) was performed by immunoprecipitation with anti-IgG antibody.

internal sequence has diverged enough such that it does not contain any E2F TFBSs, and an unrelated genomic sequence located 4 kb upstream of an E2F-TE (Figure S3).

We first analyzed the binding of E2Fa to E2F-TEs located close to genes, and have compared it to binding to known E2F target genes. Our results show that E2Fa binds to both these E2F-TEs and its known targets, but does not show binding to sequences that do not contain the E2F TFBS consensus (Figure 8 and Figure S3). The intensity of the amplified bands, and thus the abundance of the E2F/DNA complex, varied from experiment to experiment and was strongly increased at all sites in plants over-expressing the E2Fa-DPa factors (De Veylder *et al.*, 2002) (Figure 8 and Figure S3), suggesting that the available E2F factor is limiting for the binding. We detected binding to all analyzed E2F-TEs located near genes, irrespective of the epigenetic marks they are associated with, suggesting that H3K27me1-rich chromatin does not block binding of E2F to its sites.

The binding *in vivo* to E2F sites located within TEs inserted close to genes supports the hypothesis that insertion of these TEs may have a direct effect on E2F transcriptional regulation by incorporating new genes into this transcriptional network.

As already mentioned, although a proportion of the E2F-TEs are located close to genes, the rest are found far from genes, and therefore probably do not directly participate in gene promoters. We nevertheless assayed the binding of E2F to the E2F-TEs located far from genes. The ChIP analyses performed showed that the E2F-TEs located far from genes also bind E2F *in vivo* (Figure 8 and Figure S2). The binding to these sites is similar to the one previously described, showing some variability in the binding and a clear increase in plants over-expressing E2Fa-DPa factors (Figure S3).

These results show that E2F binds to all types of E2F BS, whether found in a TE or isolated, and independently of their position with respect to genes. However, this experimental system uses whole plants for the ChIP analyses, and does not allow resolution of differences between specific cells or tissues. E2F transcription factors regulate processes that are cell cycle-dependent and in some cases specific to certain organs, tissues or environmental conditions (Ramirez-Parra *et al.*, 2007). Therefore, it may be that the chromatin differences shown between isolated E2Fs and most E2F-TEs modulate the binding of E2F only in particular cells or organs or under particular environmental conditions.

CONCLUSIONS

The results presented here show that several MITEs related to at least four super-families of DNA transposons have captured and amplified a sequence that fits the consensus of the E2F BS. In most cases, the captured and amplified sequence is contained in a longer tandemly repeated unit that may be considered a mini-satellite. Our results suggest that capture of the E2F BS by these MITE families took place in an ancestral *Brassica* genome, and that the MITEs and the E2F BSs were amplified to different extents during evolution of the various *Brassica* species analyzed. As a result, in four of the five *Brassica* species analyzed, the vast majority of the E2F BSs (73% in *A. thaliana* and 85% in *A. lyrata*) are located within TEs. These E2F BSs within TEs are bound by the E2F protein *in vivo*, suggesting that they participate, directly or indirectly, in the E2F transcriptional network. The effect of E2F-TEs differs depending on their location with respect to genes. Some E2F-TEs may directly participate in gene promoters, as they are located close to genes that, in some cases, have been reported as being regulated by E2F. However, an important proportion of the E2F-TEs are located far from genes, which suggests that these elements do not directly participate in gene promoters. Nonetheless, they may still have an effect on gene regulation, and we propose two possible scenarios. On the one hand, the E2F-TEs located far from genes may constitute a reservoir of E2F BSs that may be mobilized during evolution to rewire new genes into the E2F transcriptional network. In this respect, it is interesting to note that a proportion of the E2F-TEs inserted close to genes are polymorphic among *A. thaliana* ecotypes (Table S3), suggesting that the E2F-TEs have been transpositionally active recently and that the population of genes wired into the E2F transcriptional network may be different in closely related genomes. On the other hand, the fact that the E2F-TEs located far from genes are able to bind E2F, and that these elements contain an important proportion of all E2F BSs, suggest that they may reduce the concentration of free E2F protein available to bind to the E2F BSs present in promoters. Indeed, our results show that the concentration of E2F is limiting for its binding to all E2F BSs, including those of its known targets, suggesting that amplification of the number of E2F sites has had a direct effect on E2F binding. Interestingly, the fact that E2F-TEs located far from genes are associated with local heterochromatin may suggest a mechanism to regulate their effect on the concentration of E2F. The fact that we have not been able to detect an effect of the chromatin state on the E2F binding to E2F-TEs may be due to our experimental system, which used whole plants for the ChIP analyses and does not allow resolution of differences between specific cells or tissues. E2F-TEs regulate cell cycle-associated processes, and an in-depth study of their binding regulation may require experimental approaches that allow access to individual cells or cell types. In any case, the strong influence of chromatin in transcription factor binding is well established, and it is interesting to note that the silent state of TEs, which usually correlates with their association with local heterochromatin, may be modified in response to environmental conditions. In addition, it has been reported that silencing can be relieved under particular stress conditions without any apparent change in their epigenetic marks (Bucher *et al*., 2012). Similarly, the silencing of transposons may also vary during plant development, the most striking case being their reactivation in Arabidopsis pollen (Slotkin *et al*., 2009). Therefore, modulation of the ability of the E2F-TEs to bind E2F in response to environmental or developmental signals, resulting in changes in the effective concentration of the E2F transcription factor, would constitute an additional mechanism for fine-tuning the E2F transcriptional network.

In summary, our results show that TEs have highly amplified the number of E2F BSs in various *Brassica* species, and suggest a singular way by which transposons may affect host genes by modulating essential transcriptional networks.

EXPERIMENTAL PROCEDURES

Plant material

Plants were grown on soil at 22°C under long-day conditions (16 h light/8 h dark).

Comparison of TEs in *Brassica* genomes

The genomic sequences used were the TAIR9 assembly for *A. thaliana* (www.tair.org) and the genomes available at Phytozome (www.phytozome.net) for *A. lyrata*, *C. rubella*, *T. halophila*, *B. rapa* and *Oryza sativa.*

The annotation of TEs in *C. rubella*, *T. halophila* and *B. rapa* was performed with RepeatMasker (http://www.repeatmasker.org/) using the *A. thaliana* repeat database downloaded from RepBase (www.girinst.org). MITEs were specifically searched for using SUBOTIR (J.P., unpublished data), and the predictions were merged to obtain a non-redundant annotation. We used the available TE annotation for *A. lyrata* and *A. thaliana* (version TAIR9 at www.arabidopsis.org). The families *Simplehat*1, *Simplehat*2 and *Simpleguy*1 were re-annotated in *A. thaliana* by aligning the elements of a given family as defined by TAIR9, discarding the mini-satellite region, concatenating them, and using this as a query for COPILIST (Garcia-Mas *et al*., 2012), allowing a gap up to 10 000 bp.

We based our comparisons of the TEs across genomes on sequence similarity rather than their family according to the annotations, as very divergent sequences may often be attributed to the same family using these annotation methods. To do this, we clustered all annotated TE sequences within each annotated genome using SILIX (Miele *et al*., 2011). Within each genome, we considered further only the largest clusters, which together account for over 80% of the E2F sites in that given genome. We chose the longest sequence of each cluster as its representative, and performed pairwise comparisons of all the representatives found in all the genomes, and determined two clusters as being the same family if the two representatives were 60% similar across 70% of their length.

Identification of E2F binding motifs

The coordinates of the sequences fitting the E2F BS consensus were identified using Vmatch (http://www.vmatch.de/) for perfect matches on either strand.

Identification of mini-satellites

Tandem repeats were identified using TRF (http://tandem.bu.edu/trf/trf.html) (Benson, 1999) with default parameters, except for mismatch penalty (5), indel penalty (5), minimum alignment score to report (1) and maximal length motif to report (35).

Annotation manipulations

Intersections and overlaps of sets of annotations were performed using the BedTools suite (http://code.google.com/p/bedtools/) (Quinlan and Hall, 2010).

Statistical analysis of epigenetic marks

The genome-wide epigenetic maps for eight histone variants (H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K9Ac and H3K9me2) and a positive control (H3) were obtained from the NCBI Short Read Archive database (GEO accession number GSE28398) (Luo *et al*., 2013). MACS (Zhang *et al*., 2008) was used to call peaks within this read mapping data, with default parameters and an e-value cut-off of 10^{-5}. For each epigenetic mark, we intersected the coordinates of these peaks with those of all instances of the sequence TTCCCGCCAA. We constructed a binary matrix of data points, representing all instances of this sequence in the genome. Each row describes a data point, and the columns represent the presence (1) or absence (0) of a given epigenetic mark at this position. A last column indicates whether the data point is found within (1) or outside (0) a TE. We discarded any data point that did not have any epigenetic mark as this position may not be mappable. We performed a χ^2 test between the column representing an epigenetic mark and the column of TE/non-TE labels to determine whether that mark is significantly associated with the labels or not. We then calculated the correlation, the sign of which indicates whether the association is with the TE or non-TE subset.

ChIP analyses

In vivo cross-linking and chromatin isolation from leaves of 17-day-old seedlings were performed as previously described (Bowler *et al*., 2004). Chromatin was immunoprecipitated using the antibodies anti-monomethyl histone H3 (Lys27) (H3K27me1;, Upstate Millipore, http://www.millipore.com/, reference 07–448), anti-dimethyl histone H3 (Lys4) (H3K4me2; Upstate Millipore, reference 07–030) or anti-E2Fa. For E2Fa, in addition to the standard method, immunoprecipitations using a low-cell ChIP kit (Diagenode, http://www.diagenode.com) were also performed with indistinguishable results. Immunoprecipitations using rabbit IgG (Diagenode) or no antibody were used as negative controls for immunoprecipitations performed using the low-cell kit (Diagenode) or the standard method, respectively. Immunoprecipitated DNA was analyzed by semi-quantitative PCR under standard conditions using primers amplifying individual E2F-TEs (Table S1). All ChIP experiments were performed with at least two biological replicates. The oligonucleotides used in PCR amplifications are listed in Table S1.

ACKNOWLEDGMENTS

The authors thank Joshua Baller (Center for Genome Engineering, University of Minnesota, USA) for a preliminary bioinformatic analysis of the association of E2F sites with particular epigenetic marks, Lieven De Veylder (VIB Department of Plant Systems Biology, University of Ghent, Belgium) for the E2Fa antibody, Florian Maumus (INRA, UR1164 URGI, INRA de Versailles, France) for sharing unpublished data, and Elena Casacuberta (Institut de Biologia Evolutiva, CSIC-UPF, Barcelona, Spain), M. Lluisa Espinás, (IBMB-CSIC, Barcelona, Spain) Fabien Nogué (IJPB, INRA, Versailles, France) and José Luís Riechmann (CRAG, Barcelona, Spain) for critical reading of the manuscript. This work was supported by the Ministerio de Ciencia e Innovación (grant BFU2009-11932 to J.M.C. and grants BFU2009-9783 and BFU2012-34821 to C.G.), and by an institutional grant from the Fundacion Ramon Areces to Centro de Biologia Molecular Severo Ochoa (to C.G.).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Schema summarizing the relationship between the hAT2 transposon and the *Simplehat*2 MITE family, according to the current model of MITE origin and amplification.

Figure S2. ChIP analyses of the epigenetic marks H3K4me2 and H3K27me1 associated with various types of E2F BS.

Figure S3. ChIP analyses of binding of the E2F protein to E2F BSs.

Table S1. Oligonucleotides used in PCR amplifications

Table S2. Genomic location and epigenetic context of the E2F-TEs belonging to the four major families in *A. thaliana*.

Table S3. E2F-TE insertion polymorphisms among different ecotypes of *Arabidopsis*.

REFERENCES

Ahmed, I., Sarazin, A., Bowler, C., Colot, V. and Quesneville, H. (2011) Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. *Nucleic Acids Res.* **39**, 6919–6931.

Beilstein, M., Nagalingum, N., Clements, M., Manchester, S. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **107**, 18724–18728.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Bourque, G. (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* **19**, 607–612.

Bourque, G., Leong, B., Vega, V.B. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762.

Bowler, C., Benvenuto, G., Laflamme, P., Molino, D., Probst, A.V., Tariq, M. and Paszkowski, J. (2004) Chromatin techniques for plant cells. *Plant J.* **39**, 776–789.

Bucher, E., Reinders, J. and Mirouze, M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. *Curr. Opin. Plant Biol.* **15**, 503–510.

Cowley, M. and Oakey, R.J. (2013) Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet.* **9**, e1003234.

De Veylder, L., Beeckman, T., Beemster, G.T. *et al.* (2002) Control of proliferation, endoreduplication and differentiation by the Arabidopsis E2Fa-DPa transcription factor. *EMBO J.* **21**, 1360–1368.

DeGregori, J. and Johnson, D.G. (2006) Distinct and overlapping roles for E2F family members in transcription, proliferation and apoptosis. *Curr. Mol. Med.* **6**, 739–748.

Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405.

Garcia-Mas, J., Benjak, A., Sanseverino, W. *et al.* (2012) The genome of melon (*Cucumis melo* L.). *Proc. Natl Acad. Sci. USA*, **109**, 11872–11877.

Guermonprez, H., Hénaff, E., Cifuentes, M. and Casacuberta, J. (2012) MITEs, miniature elements with a major role in plant genome evolution. In *Topics in Current Genetics: Plant Transposable Elements* (Grandbastien, M.A. and Casacuberta, J.M.,eds). Heidelberg/Berlin: Springer-Verlag, pp. 112–124.

van den Heuvel, S. and Dyson, N.J. (2008) Conserved functions of the pRB and E2F families. *Nat. Rev. Mol. Cell Biol.* **9**, 713–724.

Heyman, J., Van den Daele, H., De Wit, K., Boudolf, V., Berckmans, B., Verkest, A., Alvim Kamei, C.L., De Jaeger, G., Koncz, C. and De Veylder, L. (2011) *Arabidopsis* ULTRAVIOLET-B-INSENSITIVE4 maintains cell division activity by temporal inhibition of the anaphase-promoting complex/cyclosome. *Plant Cell*, **23**, 4394–4410.

Hu, T.T., Pattyn, P., Bakker, E.G. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.

Jacob, Y., Feng, S., LeBlanc, C.A., Bernatavichute, Y.V., Stroud, H., Cokus, S., Johnson, L.M., Pellegrini, M., Jacobsen, S.E. and Michaels, S.D. (2009) ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nat. Struct. Mol. Biol.* **16**, 763–768.

Jacques, P.-E., Jeyakani, J. and Bourque, G. (2013) The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504.

Kaul, S., Koo, H.L., Jenkins, J. *et al.* (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X.Y., Chan, Y.S., Ng, H.H. and Bourque, G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634.

Lammens, T., Li, J., Leone, G. and De Veylder, L. (2009) Atypical E2Fs: new players in the E2F transcription factor family. *Trends Cell Biol.* **19**, 111–118.

Lisch, D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61.

Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P. and Lam, E. (2013) Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant J.* **73**, 77–90.

Lynch, V.J., Leclerc, R.D., May, G. and Wagner, G.P. (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1158.

Medstrand, P., van de Lagemaat, L.N., Dunn, C.A., Landry, J.R., Svenback, D. and Mager, D.L. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet. Genome Res.* **110**, 342–352.

Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.

Naouar, N., Vandepoele, K., Lammens, T. *et al.* (2009) Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. *Plant J.* **57**, 184–194.

Ouyang, S., Zhu, W., Hamilton, J. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ramirez-Parra, E., Fründt, C. and Gutierrez, C. (2003) A genome-wide identification of E2F-regulated genes in Arabidopsis. *Plant J.* **33**, 801–811.

Ramirez-Parra, E., del Pozo, J.C., Desvoyes, B., de la Paz Sanchez, M. and Gutierrez, C. (2007) E2F–DP transcription factors. In *Annual Plant Reviews: Cell Cycle Control and Plant Development.*, Vol. 32 (D. Inze., ed.) Oxford: Blackwell Publishing Ltd, pp. 138–163.

Richard, G.F., Kerrest, A. and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* **72**, 686–727.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148** (1–2), 335–348.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J.D., Feijó, J.A. and Martienssen, R.A. (2009) Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, **136**, 461–472.

de Souza, F.S.J., Franchini, L.F. and Rubinstein, M. (2013) Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? *Mol. Biol. Evol.* **30**, 1239–1251.

Takeda, S., Sugimoto, K., Otsuki, H. and Hirochika, H. (1999) A 13-bp *cis*-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. *Plant J.* **18**, 383–393.

Testori, A., Caizzi, L., Cutrupi, S., Friard, O., De Bortoli, M., Cora, D. and Caselle, M. (2012) The role of transposable elements in shaping the combinatorial interaction of transcription factors. *BMC Genomics*, **13**, 400.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inze, D. and De Veylder, L. (2005) Genome-wide identification of potential plant E2F target genes. *Plant Physiol.* **139**, 316–328.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K. and Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA*, **104**, 18613–18618.

Wang, J.R., Bowen, N.J., Marino-Ramirez, L. and Jordan, I.K. (2009) A c-Myc regulatory subnetwork from human transposable element sequences. *Mol. BioSyst.* **5**, 1831–1839.

Zhang, Y., Liu, T., Meyer, C.A. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

Zheng, N., Fraenkel, E., Pabo, C.O. and Pavletich, N.P. (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes Dev.* **13**, 666–674.

Chapter 6
The Impact of Transposable Elements in the Evolution of Plant Genomes: From Selfish Elements to Key Players

Beatriz Contreras, Cristina Vives, Roger Castells and Josep M. Casacuberta

Abstract Transposable elements (TEs) are major components of all eukaryote genomes, and in particular of plant genomes. Whereas these elements have long been considered as selfish 'junk DNA without function', the data accumulated over the years have shown that they are essential components of the genome structure and key players of genome evolution. Here, we summarize the recent advancement in the field and we discuss the role of TEs in the light of the new data coming from whole plant genome sequences and next-generation sequencing (NGS) data on resequencing of plant varieties and lines.

6.1 Transposable Elements, a Major Component of Plant Genome

Transposable elements (TEs) are mobile genetic elements that account for an important fraction of virtually all eukaryote genomes. TEs can be classified into two major classes, class I (retrotransposons) and class II (DNA transposons). Class I elements transpose through an RNA intermediate used as a template in a reverse transcription reaction leading to a new DNA copy that can integrate back into the genome. Therefore, class I TEs do not excise during transposition and their copy number increases as a result of their movement. Whereas the transcription of the element is catalysed by the host's polymerase (Pol II), its reverse transcription and integration are catalysed by enzymatic activities encoded by the retrotransposon itself, in case of autonomous elements, or by a related element, in case of non-autonomous elements. Class II elements transpose via a DNA intermediate, which results from the excision of the element from its chromosomal location and that can be integrated elsewhere into the genome. Both the excision and integration reactions are catalysed by a transposase which is encoded by the mobilized TE in

B. Contreras · C. Vives · R. Castells · J.M. Casacuberta (✉)
Centre de Recerca En Agrigenòmica, CRAG (CSIC-IRTA-UAB-UB), Barcelona, Spain
e-mail: josep.casacuberta@cragenomica.es

P. Pontarotti (ed.), *Evolutionary Biology: Biodiversification from Genotype to Phenotype*, DOI 10.1007/978-3-319-19932-0_6

case of autonomous elements or by a related element in case of a defective TE copy. There are, however, some DNA transposons that move through a different mechanism. This is the case of *Helitrons*, which transpose via a rolling-circle mechanism similar to that of some bacterial TEs. Both class I and class II TEs can be further classified into families and subfamilies depending on their structure, encoded proteins and mechanism of transposition (Wicker et al. 2007).

Whereas TEs are commonplace in eukaryotes, and most eukaryotes contain elements belonging to all major types and classes, their prevalence differs from genome to genome. TEs account for a major but variable fraction of plant genomes (Bennetzen and Wang 2014), with LTR retrotransposons and miniature inverted-repeat transposable elements (MITEs) tending to be the most represented types of TEs (Casacuberta and Santiago 2003). The variability in TE content is huge in plants. For instance, as much as 85 % of maize genome or 70 % of Norway spruce genome (Nystedt et al. 2013) has been annotated as transposons, whereas transposon annotations make only the 21 % of the more compact *Arabidopsis thaliana* genome (Ahmed et al. 2011). These numbers are not directly comparable as the methods and the parameters used to perform the annotations are different, and this may have an important impact on the sensibility and specificity of the detection. Indeed, analyses in *A. thaliana* have shown that there is a continuum between repetitive elements and unannotated genomic dark matter, making it somehow arbitrary to define a frontier (Maumus and Quesneville 2014). However, in spite of these limitations, there seem to be a direct relationship between genome size and percentage of TEs within the genome. Analyses of closely related species, for example of the *Oryza* genus (Chénais et al. 2012), suggest that TE activity and polyploidization are the two main mechanisms responsible for genome size increase during evolution (Panaud et al. 2014). The relationship between genome duplication and transposition is interesting. On the one hand, gene duplication can allow genomes to tolerate a higher TE activity, as their mutagenic capacity is buffered by having extra copies of essential genes, but on the other hand, the lack of gene duplications may force the genome to explore other sources of innovations such as transposition. In this respect, it is interesting to note that gymnosperms, that in contrast to angiosperms do not seem to have suffered recent whole-genome duplications, present extremely big genomes with a very high content of TEs (De La Torre et al. 2014).

The effect of TE activity in genome size may be quite dramatic over short periods of time, as suggested by the high activity of TEs associated to the genome size doubling of *Oryza australiensis,* a wild relative of rice, during the last three million years (Zhao and Ma 2013). However, although TEs may be responsible for rapid genome size changes, their activity is not constant during evolution. Indeed, TEs seem to alternate periods where they are relatively quiescent with burst of transposition where their copy number increases significantly (Vitte et al. 2014). This evolutionary behaviour of transposons as a whole can be in part explained by the results obtained analysing the regulation of particular transposons and genomes. All the data accumulated so far indicate that transposons are heavily silenced in genomes by different mechanisms, and in particular by epigenetic mechanisms

(Ito and Kakutani 2014). Silent TEs of different classes, including both DNA transposons and retrotransposons, can be reactivated in mutated genetic backgrounds showing reduced DNA methylation (Ito and Kakutani 2014), which shows that the silenced TEs retain their capacity to be activated. In fact, TEs can be activated in wild-type plants in particular situations or developmental stages. TEs are de-repressed in the gametophytes and their expression may allow the production of sRNAs to ensure the maintenance of the epigenetic silencing of TEs in the following generation, although alternative explanations of this phenomenon are also possible (Martínez and Slotkin 2012). In addition, over the years, data have accumulated on the stress-related activation of different TEs. This includes the well-studied activation of the tobacco retrotransposon *Tnt1* by biotic and abiotic stresses (Grandbastien et al. 2005), the cold and salt activation of the rice MITE *mPing* (Naito et al. 2009) and the heat activation of the Arabidopsis *ONSEN* retrotransposons (Cavrak et al. 2014). Similarly, it is known that in vitro culture, which can be considered as a complex stress, can reactivate TEs in rice and maize (Hirochika 1997; Kaeppler et al. 2000). Plants are subjected to stress in nature, and this may lead to reactivation of TEs in certain cells. In most cases, the somatic activation of TEs will not lead to germinal transpositions and therefore will not be inherited by the successive generations. However, in particular situations, a general release of the control mechanism may lead to a general activation of TEs leading to a burst of transposition. It is interesting to note that it has been shown that interspecific crosses or polyploidization events may lead to global epigenetic changes and activation of TEs (Parisod et al. 2009; Yaakov and Kashkush 2011). As these phenomena are commonplace in plant evolution, this may give the opportunity to TE amplification bursts to occur and accompany speciation events.

6.2 Transposable Elements in Genome Structure

TEs are usually not homogeneously distributed along chromosomes. They concentrate in pericentromeric regions, while they are less abundant in chromosome arms, in a pattern that is usually complementary to that of genes. These pattern of TEs can be the consequence of both a preferential insertion into these regions, as demonstrated for yeast retroelements, or the effect of selection cleaning up the more frequently deleterious TE insertions in gene-rich regions (Neumann et al. 2011; Peterson-Burch et al. 2004). Selection against insertion within genes, which are not homogeneously distributed along chromosomes, and the recombination rate, which is also different in different chromosomal regions and greatly influences TE elimination, explains in part the distribution of TEs (Bennetzen and Wang 2014). However, it has been shown that some TEs indeed have a preferential insertion into certain genomic regions. In general, *Copia*-like TEs show some preference for gene-rich regions, whereas *Gypsy*-like TEs are supposed to target preferentially the heterochromatic pericentromeric regions (Peterson-Burch et al. 2004). As an example, the tobacco *Tnt1* and the rice *Tos17 Copia* elements preferentially insert into gene-rich regions (Miyao et al. 2003; Le et al. 2007), whereas in cereals, there are some families of *Gypsy* retrotransposons that are almost exclusively located in the centromeres, suggesting a high preference for insertion into these regions (Gao et al. 2009; Wolfgruber et al. 2009; Langdon et al. 2000; Li et al. 2013; Jiang et al. 2003). However, there are exceptions to this rule, and some *Gypsy* elements such as the low-copy-number *LORE1* retrotransposon from *Lotus japonicus* seem to target gene-rich regions (Madsen et al. 2005) and some *Copia*-like retrotransposons such as the *Tal1* element from *Arabidopsis lyrata* target the centromere for integration (Tsukahara et al. 2012).

The fact that TEs, and in particular high-copy-number retrotransposons, tend to concentrate in gene-poor heterochromatic regions, does not imply that they do not impact on genome function. Indeed, TE insertions in the pericentromeric regions probably have a profound impact on the structure and dynamics of genomes. The main mechanism to control the activity of TEs is their epigenetic silencing. As a consequence of their silencing, TE sequences tend to be heavily methylated and are associated with expression-repressive histone modifications (Ito and Kakutani 2014). Therefore, the concentration of TEs in the centromere also concentrates certain epigenetic marks in these regions, leading to a particular chromatin structure that is essential for heterochromatin compaction and function in the centromeres (Wong and Choo 2004). It has been proposed that TEs, and in particular LTR retrotransposons sitting in the centromere, may transcribe flanking repeats and other centromeric sequences leading to the production of double-stranded RNA which would direct their particular heterochromatic structure (Lippman et al. 2004). In fact, studies on the formation of neocentromeres have shown that it is the epigenetic nature of centromere elements, and not their sequence, which ensures its functionality (Zhang et al. 2013). Therefore, there is probably a dynamic interplay between retrotransposons and heterochromatin where some TEs target heterochromatin for integration (in the case of *Gypsy*-like elements through the chromodomains of their integrases that are known to interact with some heterochromatic epigenetic marks) and help thereafter to maintain heterochromatin by directing their epigenetic modification (Gao et al. 2008).

6.3 Transposable Elements as a Source of New Functions

TEs impact on genome and gene evolution in many ways. Perhaps, the most obvious is the generation of null mutations by transposing into a gene. Some of these null mutations have been selected by humans during plant domestication such as the waxy and sticky varieties of foxtail millet (*Setaria italica*), or Mendel's wrinkled peas (Lisch 2013). For TEs that transpose by a cut-and-paste mechanism (e.g. most class II TEs), the excision of the element may result in function recovery giving rise to mosaic phenotypes as exemplified by the kernel colour of maize cobs. Nevertheless, in some cases the excision may leave behind parts of the element that are not removed and can modify the coding capacity of the gene, and in some cases provide new gene functions (Lisch 2013; Oliver et al. 2013). This process by which a TE, or a part of it, is established in a specific region and gains a cellular function is known as molecular domestication (Kajihara et al. 2012).

There is an important number of plant genes with a transposon origin (Oliver et al. 2013; Bennetzen and Wang 2014). In particular, several important transcription factors derive from class II transposases. For example, *Daysleeper*, a transcription factor that regulates the morphogenetic development in *A. thaliana,* is derived from a *hAT* transposase (Bundock and Hooykaas 2005), or the light response FHY3 and FAR1 transcription factors that are ancient *Mutator* transposases (Hudson et al. 2003; Lin et al. 2007).

Transposons can also capture, duplicate and mobilize genes or gene fragments, creating new opportunities for gene evolution. Retrotransposons duplicate host genes or gene fragments through the reverse transcription of their mRNAs generating what is called a retrogene. The retroposed gene fragments can be fused to host genes to generate new chimeric proteins (Elrouby and Bureau 2010), and retroposed retrogenes can be regulated differently to the original genes (Abdelsamad and Pecinka 2014), which can be a source of gene innovation. Class II transposons can also transduplicate genes. Pack-MULEs, for example, are *Mutator*-like TEs that carry fragments of genes in different plants and were proposed as important mediators of gene evolution in plants (Jiang et al. 2004). The fact that an important fraction of rice Pack-MULEs is transcribed and show signs of purifying selection suggested that indeed these elements have a role in gene evolution in plants (Hanada et al. 2009). A part from MULEs, other class II TEs, such as *CACTA* elements, have been shown to transduplicate host gene fragments in different plants (Benjak et al. 2008; Morgante 2006). But probably the TEs that seem to capture more actively, amplify and mobilize gene fragments are the rolling-circle transposing elements *Helitrons*. More than one-third of the thousands *Helitrons* of maize genome carry at least one host gene fragment (Du et al. 2009). Therefore, TEs have a great potential to generate new gene structures by shuffling host genome sequences (Bennetzen 2005; Morgante 2006).

6.4 Impact of Transposable Elements in Gene Regulation

In addition to their effect on the coding capacity of the host genome, TEs can impact on host genes in many ways. As already explained, the expression of TEs is tightly regulated, both because they are the main target of the silencing machinery and also because they usually have stress-related promoters that are only active under particular situations. For this reason, in addition to being able to modify host gene expression by interrupting gene regulatory regions upon insertion, for example in the case of the *Vgt1* regulatory locus of maize (Salvi et al. 2007), TEs can modify the expression of host genes located nearby by contributing their own regulatory elements or by attracting the silencing machinery.

There are several examples of insertions of TEs that induce new transcriptional regulations to host genes. This is the case of the insertion of a *Hopscotch* TE some 50 Kb upstream of the *theosinte branched* 1 (*tb1*) gene, which represses branching in maize, which results in its overexpression and the apical dominant phenotype of modern maize (Studer et al. 2011) or the insertion of an LTR retrotransposon upstream of the *Ruby* gene in oranges which confers to this gene a developmental regulation and cold inducibility resulting in the blood orange phenotype (Butelli et al. 2012).

MITEs are a particular type of transposons present in high copy numbers in plant genomes (Casacuberta and Santiago 2003). They are relatively small, which may help them avoiding to generate complete knockout phenotypes, and although they do not need to be expressed to transpose, they can contain transcriptional regulatory sequences. For example the rice *mPing* MITE contains stress-responsible transcriptional regulatory elements that upregulate neighbouring genes under cold and salt stress conditions (Yasuda et al. 2013; Naito et al. 2009). The high copy number of MITEs makes them particularly suited to modify the expression of groups of genes, making it possible to create, or to extend, transcriptional regulatory networks. The fact that some transcription factors derive from transposases (see above), and that the sequences bound by transposases (e.g. the TIRs) can be mobilized throughout the genome, was proposed as a potential mechanisms to create and modify transcriptional regulatory networks (Feschotte 2008). In the recent years, evidences that TEs can mobilize transcription factor binding sites and rewire transcriptional networks have accumulated (Rebollo et al. 2012). In plants, a recent report from our laboratory has shown that different families of MITEs have amplified and redistributed the binding sites for the E2F transcription factor during *Brassica* evolution, and the insertion of some of these MITEs may have wired new genes into the E2F transcriptional network (Hénaff et al. 2014).

In spite of the examples explained above that illustrate the potential of TEs to bring new regulatory sequences to host genes, the most frequent effect of a TE insertion within or close a gene promoter is its inactivation. As already explained, TEs are controlled by epigenetic mechanisms that silence them tightly. For this reason, most TEs are heavily methylated and are associated to inactive chromatin, and this can influence genes located nearby that can become silenced by the presence of the TE. A well-studied example of such an effect is the epigenetic silencing of a sex determination gene in melon linked to a TE insertion in its upstream region (Martin et al. 2009). Similarly, the necessary repression of the flowering regulator *FWA* gene in *A. thaliana* is a consequence of the epigenetic silencing of a SINE transposon located in its promoter (Kinoshita et al. 2007). Genome-wide analyses suggest that these effects may be highly relevant. As an example, it has been shown that about 300 genes differentially expressed in maize populations have changes in DNA methylation, and many of these regions are associated with transposons (Eichten et al. 2013). This suggests that polymorphic TE insertions modify the pattern of genome methylation which translates into changes in gene expression.

Silencing of TEs is mediated by siRNAs that target TE sequences which probably originate from the expression of particular TE structures (e.g. inverse repeated elements). Whereas the main target of these siRNAs are TEs, in some cases TEs may produce siRNAs that target host genes (Bennetzen and Wang 2014; McCue and Slotkin 2012). In fact, it has been proposed that TE can be the source of both siRNAs and miRNAs (Li et al. 2011; Piriyapongsa and Jordan 2008) which suggests that the genome has evolved a new layer of gene regulation from its defence mechanisms against TEs.

The expression of TEs may also interfere with host genes creating sense or antisense transcripts that may result in their specific silencing. It has been shown that read-through transcription, due to a leaky transcriptional terminator, is relatively frequent in plant retrotransposons, and this could result in the inclusion of flanking sequences into retrotransposon transcripts. As a consequence, as it has been shown in tobacco (Hernández-Pinzón et al. 2009), the convergent transcription of a retrotransposon located downstream of a host gene could result in the formation of dsRNAs which may potentially regulate the host gene. In addition, TEs insertions in 5' leader region, 3' trailer sequence or introns can modify the sites of RNA processing or polyadenylation affecting gene expression (Bennetzen and Wang 2014).

6.5 Transposable Elements Dynamics and Evolution of Crop Plants

We have seen in the previous sections that TEs can impact on genomes in many ways, from providing new genes or modifying the existing ones or alter their expression, to modify genome or chromosome structure. Because of that TEs are an extraordinary source of novelty useful for evolution (Lisch 2013). In particular, in the last few years, a number of examples of TE insertions leading to important agronomic traits that have been selected during evolution and breeding have accumulated (Lisch 2013). These include the different flesh fruit colour in blood orange (Butelli et al. 2012), the different skin colours in grapevine (This et al. 2007), the nectarine phenotype in peaches (Vendramin et al. 2014) or the seedless phenotype in apples (Yao et al. 2001) (see Fig. 6.1). However, evaluating the impact of TEs in the evolution of eukaryote genomes is not an easy task. In spite of the examples listed above on TEs that gave rise to mutations that have been selected during evolution, a general evaluation is still lacking. There are several reasons for that, as previously pointed out (Vitte et al. 2014). Although the number of plant genomes sequenced is growing rapidly, the quality of the published genomes is not always good enough to allow a proper analysis of the TE content. Indeed, most published genomes contain a variable, and usually important, fraction of unassembled reads which are usually enriched in repetitive sequences including TEs. This precludes a complete genome-wide TE analysis. In addition to the quality of

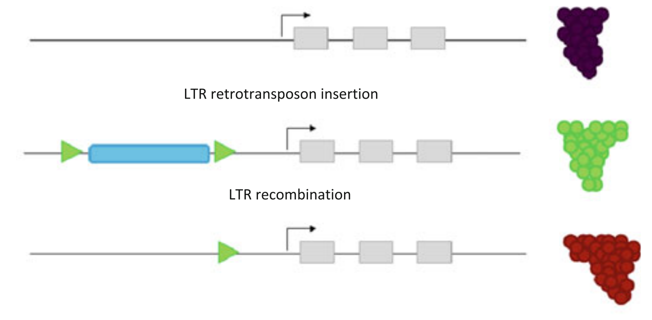
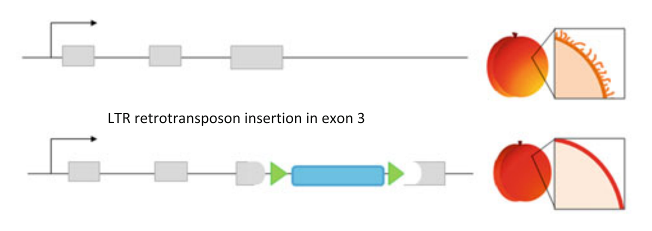
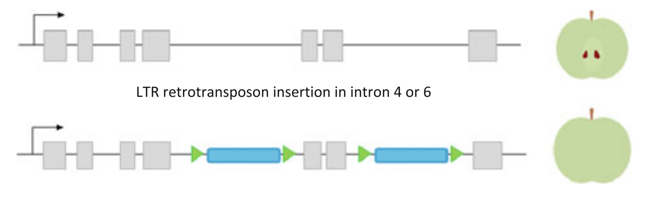
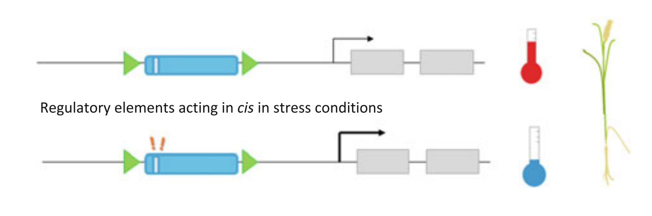
TE related traits in crops	Representation
TE insertion in promoter sequence confers different grape skin colours. (This et al. 2007)	
Peach and nectarine skin phenotype due to a *Copia* retrotransposon insertion in exon sequence. (Vendramin et al. 2014)	
Seedless phenotype in some apple varieties is due to a retrotransposon insertion in an intron. (Yao et al. 2001)	
A TE insertion in regulatory region confers cold specificity regulation in rice. (Yasuda et al. 2013)	

Fig. 6.1 Representation of different important agronomic traits that are due to transposable element insertions. *Grey boxes* represent exons, *blue boxes* represent TE coding region, and *green triangles* represent LTRs

the sequence and assembly, the annotation of the TE content is also highly variable among the sequenced genomes. There are several reasons for that, including the use of different bioinformatics tools and pipelines as well as the thresholds set which determine the sensitivity and specificity of the annotation tools. This makes comparisons of the TE content between genomes a very difficult exercise, and different voices claim that there is a need for an international effort to standardize the methods used for annotating TEs (Hoen, Bureau, Bourke and Blanchette, in preparation). But even with good genome sequences and TE annotation, reference genomes are only a snapshot, a fixed image, of a genome and analysing the impact of TEs in genome evolution will require sequence variability analysis within a species or among different related species. In the last few years, an important amount of resequencing data of crop varieties and landraces has being accumulated. As an example, 3000 rice varieties have already been sequenced and offer an unprecedented opportunity to search for the genetic bases of a wide range of phenotypic differences (Li et al. 2014). However, in most cases, the analyses of variability are restricted to SNPs, and TE insertion polymorphisms are not analysed. The reason for that is that detecting TE polymorphisms, and in particular TE insertions with respect to the reference genome is far from trivial. There are a number of recent tools that allow detecting TE insertion polymorphisms using paired-end resequencing data, including TEA (Lee et al. 2012), RetroSeq (Keane et al. 2013), VariationHunter (Hormozdiari et al. 2010), TEMP (Zhuang et al. 2014) and Jitterbug (Hénaff et al. submitted), but they are only starting to be used to determine the role of TEs in plant genome evolution (see for example Sanseverino et al., submitted). The use of these tools on the growing amount of resequencing data on plant varieties and accessions will probably allow us in the next future to have a more global and complete view of the impact of TEs in plant genome evolution. In particular, the analysis of crop genomes and the comparison of crop reference genomes with that of, on the one hand, their wild ancestors, and on the other hand, domesticated landraces or elite varieties will shed light on the role of TEs on the evolution of plant genomes during domestication and breeding. In addition, as crop domestication is an excellent model to study genome evolution at large, as it has already been said (Olsen and Wendel 2013), these analyses will probably allow us to better understand the structure and evolution of plant genomes and the key role played by TEs, who once were called junk DNA and now are rediscovered as key factors for genetic innovation.

References

Abdelsamad A, Pecinka A (2014) Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming. Plant Cell 26:3299–3313. doi:10.1105/tpc.114.126011

Ahmed I, Sarazin A, Bowler C et al (2011) Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis. Nucleic Acids Res 39:6919–6931. doi:10.1093/nar/gkr324

Benjak A, Forneck A, Casacuberta JM (2008) Genome-wide analysis of the “cut-and-paste” transposons of grapevine. PLoS ONE 3:e3107. doi:10.1371/journal.pone.0003107

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627. doi:10.1016/j.gde.2005.09.010

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65:505–530. doi:10.1146/annurev-arplant-050213-035811

Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. Nature 436:282–284. doi:10.1038/nature03667

Butelli E, Licciardello C, Zhang Y et al (2012) Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24:1242–1255. doi:10.1105/tpc.111.095232

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1–11. doi:10.1016/S0378-1119(03)00557-2

Cavrak VV, Lettner N, Jamge S et al (2014) How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet 10:e1004115. doi:10.1371/journal.pgen.1004115

Chénais B, Caruso A, Hiard S, Casse N (2012) The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene 509:7–15. doi:10.1016/j.gene.2012.07.042

De La Torre AR, Birol I, Bousquet J et al (2014) Insights into conifer giga-genomes. Plant Physiol 166:1724–1732. doi:10.1104/pp.114.248708

Du C, Fefelova N, Caronna J et al (2009) The polychromatic Helitron landscape of the maize genome. Proc Natl Acad Sci U S A 106:19916–19921. doi:10.1073/pnas.0904742106

Eichten SR, Briskine R, Song J et al (2013) Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell 25:2783–2797. doi:10.1105/tpc.113.114793

Elrouby N, Bureau TE (2010) Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol 153:1413–1424. doi:10.1104/pp.110.157420

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405. doi:10.1038/nrg2337

Gao D, Gill N, Kim H-R et al (2009) A lineage-specific centromere retrotransposon in Oryza brachyantha. Plant J 60:820–831. doi:10.1111/j.1365-313X.2009.04005.x

Gao X, Hou Y, Ebina H, et al (2008) Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 359–369. doi:10.1101/gr.7146408.1

Grandbastien M, Audeon C, Bonnivard E et al (2005) Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res 110:229–241. doi:10.1159/000084957

Hanada K, Vallejo V, Nobuta K et al (2009) The functional role of pack-MULEs in rice inferred from purifying selection and expression profile. Plant Cell 21:25–38. doi:10.1105/tpc.108.063206

Hénaff E, Vives C, Desvoyes B et al (2014) Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. Plant J 77:852–862. doi:10.1111/tpj.12434

Hénaff E, Zapata L, Casacuberta JM, Ossowski S. (Submitted) Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution

Hernández-Pinzón I, de Jesús E, Santiago N, Casacuberta JM (2009) The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. J Mol Evol 68:269–278. doi:10.1007/s00239-009-9204-y

Hirochika H (1997) Retrotransposons of rice: their regulation and use for genome analysis. Plant Mol Biol 35:231–240

Hormozdiari F, Hajirasouliha I, Dao P et al (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26:i350–i357. doi:10.1093/bioinformatics/btq216

Hudson M, Lisch D, Quail P (2003) The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 453–471

Ito H, Kakutani T (2014) Control of transposable elements in Arabidopsis thaliana. Chromosome Res 22:217–223. doi:10.1007/s10577-014-9417-9

Jiang J, Birchler J a, Parrott W a, Kelly Dawe R (2003) A molecular view of plant centromeres. Trends Plant Sci 8:570–575. doi:10.1016/j.tplants.2003.10.011

Jiang N, Bao Z, Zhang X et al (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573. doi:10.1038/nature02945.1

Kaeppler S, Kaeppler H, Rhee Y (2000) Epigenetic aspects of somaclonal variation in plants. Plant Mol Biol 179–188

Kajihara D, Godoy F, Hamaji T et al (2012) Functional characterization of sugarcane mustang domesticated transposases and comparative diversity in sugarcane, rice, maize and sorghum. Mol Biol 639:632–639. doi:10.1590/S1415-47572012005000038

Keane TM, Wong K, Adams DJ (2013) RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29:389–390. doi:10.1093/bioinformatics/bts697

Kinoshita Y, Saze H, Kinoshita T et al (2007) Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. Plant J 49:38–45. doi:10.1111/j.1365-313X.2006.02936.x

Langdon T, Seago C, Mende M et al (2000) Retrotransposon evolution in diverse plant genomes. 156(1):313–325

Le QH, Melayah D, Bonnivard E et al (2007) Distribution dynamics of the Tnt1 retrotransposon in tobacco. Mol Genet Genomics 278:639–651. doi:10.1007/s00438-007-0281-6

Lee E, Iskow R, Yang L, Gokcumen O (2012) Landscape of somatic retrotransposition in human cancers. Science 337:967–971. doi:10.1126/science.1222077.Landscape

Li B, Choulet F, Heng Y et al (2013) Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. Plant J 73:952–965. doi:10.1111/tpj.12086

Li J-Y, Wang J, Zeigler RS (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience 3:8. doi:10.1186/2047-217X-3-8

Li Y, Li C, Xia J, Jin Y (2011) Domestication of transposable elements into MicroRNA genes in plants. PLoS ONE 6:e19212. doi:10.1371/journal.pone.0019212

Lin R, Ding L, Casola C, Ripoll D (2007) Transposase-derived transcription factors regulate light signaling in Arabidopsis. Science 318:1302–1305

Lippman Z, Gendrel A, Black M (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature. doi:10.1038/nature02724.1

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14:49–61. doi:10.1038/nrg3374

Madsen LH, Fukai E, Radutoiu S et al (2005) LORE1, an active low-copy-number TY3-gypsy retrotransposon family in the model legume Lotus japonicus. Plant J 44:372–381. doi:10.1111/j.1365-313X.2005.02534.x

Martin A, Troadec C, Boualem A et al (2009) A transposon-induced epigenetic change leads to sex determination in melon. Nature 461:1135–1138. doi:10.1038/nature08498

Martínez G, Slotkin RK (2012) Developmental relaxation of transposable element silencing in plants: functional or byproduct? Curr Opin Plant Biol 15:496–502. doi:10.1016/j.pbi.2012.09.001

Maumus F, Quesneville H (2014) Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. PLoS ONE 9:e94101. doi:10.1371/journal.pone.0094101

McCue AD, Slotkin RK (2012) Transposable element small RNAs as regulators of gene expression. Trends Genet 28:616–623. doi:10.1016/j.tig.2012.09.001

Miyao A, Tanaka K, Murata K et al (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell 15:1771–1780. doi:10.1105/tpc.012559.ements

Morgante M (2006) Plant genome organisation and diversity: the year of the junk! Curr Opin Biotechnol 17:168–173. doi:10.1016/j.copbio.2006.03.001

Naito K, Zhang F, Tsukiyama T et al (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461:1130–1134. doi:10.1038/nature08479

Neumann P, Navrátilová A, Koblížková A et al (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2:4. doi:10.1186/1759-8753-2-4

Nystedt B, Street NR, Wetterbom A et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584. doi:10.1038/nature12211

Oliver KR, McComb JA, Greene WK (2013) Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol Evol 5:1886–1901. doi:10.1093/gbe/evt141

Olsen KM, Wendel JF (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol 64:47–70. doi:10.1146/annurev-arplant-050312-120048

Panaud O, Jackson S, Wendel J (2014) Drivers and dynamics of diversity in plant genomes. New Phytol 202:15–18. doi:10.1111/nph.12633

Parisod C, Salmon A, Zerjal T et al (2009) Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol 184 (4):1003–1015. doi:10.1111/j.1469-8137.2009.03029.x

Peterson-Burch B, Nettleton D (2004) Voytas D (2004) Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome Biol 5:R78

Piriyapongsa J, Jordan I (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. Rna 814–821. doi:10.1261/rna.916708.ferred

Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46:21–42. doi:10.1146/annurev-genet-110711-155621

Salvi S, Sponza G, Morgante M et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A 104:11376–11381

Sanseverino W, Hénaff E, Vives C, et al. (submitted) The contribution of transposon insertion polymorphisms and nucleotide variability to the evolution of the melon genome

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet 43:1160–1163. doi:10.1038/ng.942.Identification

This P, Lacombe T, Cadle-Davidson M, Owens CL (2007) Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1. Theor Appl Genet 114:723–730. doi:10.1007/s00122-006-0472-2

Tsukahara S, Kawabe A, Kobayashi A (2012) Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes Dev 26:705–713. doi:10.1101/gad.183871.111 Epub 2012 Mar 19

Vendramin E, Pea G, Dondini L et al (2014) A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach. PLoS ONE 9:e90574. doi:10.1371/journal.pone.0090574

Vitte C, Fustier M-A, Alix K, Tenaillon MI (2014) The bright side of transposons in crop evolution. Brief Funct Genomics 13:276–295. doi:10.1093/bfgp/elu002

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

Wolfgruber TK, Sharma A, Schneider KL et al (2009) Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet 5:e1000743. doi:10.1371/journal.pgen.1000743

Wong LH, Choo KHA (2004) Evolutionary dynamics of transposable elements at the centromere. Trends Genet 20:611–616. doi:10.1016/j.tig.2004.09.011

Yaakov B, Kashkush K (2011) Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. Genome 54:42–49. doi:10.1139/G10-091

Yao J, Dong Y, Morris B (2001) Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. Proc Natl Acad Sci U S A 98:1306–1311

Yasuda K, Ito M, Sugita T et al (2013) Utilization of transposable element mPing as a novel genetic tool for modification of the stress response in rice. Mol Breed 32:505–516. doi:10.1007/s11032-013-9885-1

Zhang B, Lv Z, Pang J et al (2013) Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences. Plant Cell 25:1979–1989. doi:10.1105/tpc.113.110015

Zhao M, Ma J (2013) Co-evolution of plant LTR-retrotransposons and their host genomes. Protein Cell 4:493–501. doi:10.1007/s13238-013-3037-6

Zhuang J, Wang J, Theurkauf W, Weng Z (2014) TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res 42:6826–6838. doi:10.1093/nar/gku323

MBE Advance Access published August 3, 2015

Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome

Walter Sanseverino,[†,‡,1] Elizabeth Hénaff,[†,§,2] Cristina Vives,[2] Sara Pinosio,[3] William Burgos-Paz,[¶,2] Michele Morgante,[3] Sebastián E. Ramos-Onsins,[*,2] Jordi Garcia-Mas,[*,1] and Josep Maria Casacuberta[*,2]

[1]Institut de Recerca i Tecnologia Agroalimentàries, Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Barcelona, Spain

[2]Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Barcelona, Spain

[3]Dipartimento di szience agrarie e ambientali, Università degli studi di Udine, Udine, Italy

[†]These authors contributed equally to this work.

[‡]Present address: Sequentia Biotech, Campus UAB, Edifici CRAG, Bellaterra, Cerdanyola del Vallès, Barcelona, Spain

[§]Present address: Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY

[¶]Present address: Programa de Mejoramiento Genético, Universidad de Nariño. Ciudadela Universitaria Torobajo, Pasto, Colombia

***Corresponding author:** E-mail: sebastian.ramos@cragenomica.es; jordi.garcia@irta.cat; josep.casacuberta@cragenomica.es.

Associate editor: Michael Purugganan

Abstract

The availability of extensive databases of crop genome sequences should allow analysis of crop variability at an unprecedented scale, which should have an important impact in plant breeding. However, up to now the analysis of genetic variability at the whole-genome scale has been mainly restricted to single nucleotide polymorphisms (SNPs). This is a strong limitation as structural variation (SV) and transposon insertion polymorphisms are frequent in plant species and have had an important mutational role in crop domestication and breeding. Here, we present the first comprehensive analysis of melon genetic diversity, which includes a detailed analysis of SNPs, SV, and transposon insertion polymorphisms. The variability found among seven melon varieties representing the species diversity and including wild accessions and highly breed lines, is relatively high due in part to the marked divergence of some lineages. The diversity is distributed nonuniformly across the genome, being lower at the extremes of the chromosomes and higher in the pericentromeric regions, which is compatible with the effect of purifying selection and recombination forces over functional regions. Additionally, this variability is greatly reduced among elite varieties, probably due to selection during breeding. We have found some chromosomal regions showing a high differentiation of the elite varieties versus the rest, which could be considered as strongly selected candidate regions. Our data also suggest that transposons and SV may be at the origin of an important fraction of the variability in melon, which highlights the importance of analyzing all types of genetic variability to understand crop genome evolution.

Key words: **transposon polymorphism, SNP, structural variation, melon, evolution.**

Introduction

Improving cultivars by breeding is essential to increase plant yield and ensure food security. While traditional breeding and marker-assisted breeding have been extremely successful in the past, the challenges agriculture has to face, which include the need to feed a growing human population, scarcity of land and water available for agriculture, and climate change that will drastically modify growth conditions, impose an urgent need for improving these techniques (Godfray et al. 2010). The availability of huge databases of crop genome sequences promises a new leap in plant breeding. However, in order to bridge the gap between genome sequences and trait improvement, there is a need to understand the links between genome variability and phenotypic variation. Resequencing crop varieties with interesting phenotypic traits to analyze their genomic variability promises to be an exceptional approach (Gebhardt 2013; Huang et al. 2013; Myles 2013). The analysis of genome variability using resequencing data has been used to shed light on the domestication history of different crops including, for example, maize (Hufford et al. 2012; Jiao et al. 2012), rice (Xu et al. 2012), peach (Verde et al. 2013), cucumber (Qi et al. 2013), soybean (Lam et al. 2010), watermelon (Guo et al. 2013), and tomato (Lin et al. 2014). Genomic regions with low genetic variability among domesticated varieties have been detected by these approaches, which likely highlight the genes that were selected for during domestication. In a similar way, the comparison of whole-genome sequences of varieties with contrasting phenotypic traits can be used as a means to identify genes responsible for particular agronomic traits. Sequencing the genome of parental lines and high-resolution genotyping of recombinant inbred lines identified candidate genes for quantitative trait loci associated to the increased yield of hybrid rice varieties (Gao et al. 2013). In summary, the analysis of genetic variability at the whole-genome level among varieties and cultivars should enable a new approach in plant breeding that will strengthen currently used strategies such as genome-wide association analyses as it has been recently proposed for rice (Huang et al. 2013).

The vast majority of published studies dealing with genetic variability at the whole-genome level in plants have concentrated in single nucleotide polymorphisms (SNPs) as the main type of genetic variability (3KRGP 2014; Lin et al. 2014; Qi et al. 2013). However, genomes are rife with other types of variations, and many other types of sequence modifications are responsible for genetic variability relevant for plant genome evolution. Structural variation (SV), including copy number variation (CNV) and presence/absence variation (PAV), has been shown to be frequent in plant species (Saxena et al. 2014). These SVs have traditionally been discovered using microarray-based methods, but the advent of next-generation sequencing technologies has made it possible to do so in a nonbiased manner, although the methods to do so remain computationally challenging. A well analyzed example is maize, where two inbred lines may differ by more than 50% of the genome due in part to very frequent PAV of genes (Brunner et al. 2005; Morgante et al. 2007; Springer et al. 2009). This variability in the genic component within individuals of the same species justifies the introduction of the concept of the pan-genome to refer to the ensemble of the core set of genes common to all individuals and the dispensable genome fraction specific to some of them. Interestingly, this high genome variability is enriched at loci associated with important traits (Chia et al. 2012), and translates into transcriptome differences. Indeed, a recent transcriptome analysis of 503 maize inbred lines has shown that only the 16% of transcripts are present in all lines whereas the remaining 83% are expressed in subsets of the lines (Hirsch et al. 2014). Therefore, this variability can be at the origin of phenotypic diversity of traits important for fitness and adaptation (Olsen and Wendel 2013; Hirsch et al. 2014) and therefore of agricultural importance.

Transposons are at the origin of an important fraction of the CNV and PAV due to their capacity of mobilizing gene sequences within the genome (Morgante et al. 2007). They can also contribute to genetic diversity in many other ways, the most important being the generation of transposon insertion polymorphisms. Indeed, transposon-related polymorphisms are at the origin of an important fraction of variability relevant for plant genome evolution both in the wild and in breeding processes (Lisch 2013; Olsen and Wendel 2013). For example, it has been shown that the critical increase in expression of the maize domestication gene *tb1* was the consequence of a transposon insertion in its promoter (Studer et al. 2011). Similarly, accumulated evidence shows that transposon insertion polymorphisms are responsible for phenotypic variation in agronomically important traits such as the skin or flesh color of the orange, grape, and peach (Kobayashi et al. 2004; Butelli et al. 2012; Falchi et al. 2013). A recent survey of 60 genes related to plant domestication and breeding showed that 15% of them harbor transposable element (TE) insertions that have functional effects, which suggests that TEs have an important mutational role in domesticated plant genomes (Meyer and Purugganan 2013).

In addition to genetic variation, epigenetic variation has also been shown to be highly relevant for plant evolution and transposons can be major mediators of such variability (Lisch 2013; Pecinka et al. 2013). Analyses of maize populations reveal that changes in DNA methylation are associated with changes in expression of some 300 genes, and that many of these differentially methylated regions are associated with transposons (Eichten et al. 2013). Transposons are also at the origin of variations in the epigenetic state of genes responsible for important agronomic traits. For example, changes in sex determination in melon are due to the epigenetic silencing of a sex determination gene induced by an upstream transposon insertion (Martin et al. 2009). It is thus highly relevant to study epigenetic variability in crops and to pay particular attention to transposon polymorphisms.

Melon (*Cucumis melo* L., $2n = 2x = 24$) is an important vegetable crop of the Cucurbitaceae family that is highly appreciated for its fruit quality. It has been proposed that melon originated in Africa, although recent studies suggest a possible Asian origin (Sebastian et al. 2010). It is a highly diverse species that has been classified in two subspecies, *melo* and *agrestis* (Jeffrey 1980), according to the pubescence of the female flower hypanthium although it has been shown that this classification does not completely agree with the molecular phylogeny (Stepansky et al. 1999). Both subspecies have been further divided in several botanical groups, which include both edible and wild varieties (Pitrat 2008). Genetic diversity in melon has been studied using several types of molecular markers, such as restriction fragment length polymorphism (Silberstein et al. 1999), random amplified polymorphic DNA (Stepansky et al. 1999), amplified fragment length polymorphism (Garcia-Mas et al. 2000), simple sequence repeat (Monforte et al. 2003), and SNPs (Esteras et al. 2013). The reference genome sequence of melon is available for DHL92 (Garcia-Mas et al. 2012), a double-haploid line derived from the cross between PI 161375 (Songwhan charmi [SC]) (*conomon* group, ssp. *agrestis*) and the “Piel de sapo” line T111 (PS) (*inodorus* group, ssp. *melo*). The 375 Mb assembled melon genome contains 27,427 annotated genes, and 19.7 % of the sequence was shown to correspond to TEs (Garcia-Mas et al. 2012). However, this was a conservative annotation of the most recent TEs. A less-conservative annotation showed that up to 40% of melon genome is composed of TE-related sequences (unpublished).

Here, we present the first analysis of melon genetic diversity at the whole-genome level using resequencing data from seven melon accessions from diverse origins, which includes a detailed analysis of SNPs, SV (including PAV, CNV, and inversions), and transposon insertion polymorphisms. To this end we used an array of already available and newly developed bioinformatic tools.

Results and Discussion

SNP Identification from Resequence Data

Three and four melon accessions of the ssp. *melo* and the ssp. *agrestis* subspecies, respectively, were selected as representative of the main melon groups (supplementary table

S1, Supplementary Material online). Some of these accessions are parental lines for several melon mapping populations which have been extensively used for constructing genetic maps and mapping agronomically important traits. The DHL92 line, which is the genotype of the published melon reference genome (Garcia-Mas et al. 2012) was also included in the analysis as a control. The homozygous DHL92 double-haploid line is derived from the cross between two varieties also included in this study, PI 161375 (Songwhan Charmi, spp. *agrestis*) (SC) and the Piel de Sapo T111 line (ssp. *melo*) (PS).

The seven melon varieties were resequenced using a paired-end approach, with libraries of 500 bp fragment length and 150 bp reads sequenced to an average of 19.45× depth and 80.3% breadth coverage of the assembled genome for each line (supplementary table S2, Supplementary Material online). In total we produced 273 million paired-end reads (63.9 Gb), which were mapped to the reference genome DHL92 v3.5 (Garcia-Mas et al. 2012) and variants were called using the SUPERW pipeline (supplementary fig. S1, Supplementary Material online). A total of 4,556,377 SNPs and 718,832 short deletion and insertion polymorphisms (DIPs <200 bp) were identified (table 1). A high proportion of SNPs between PS and SC have been validated using the GoldenGate and Fluidigm platforms in the context of other projects (Argyris et al. 2015).

Nucleotide Diversity at the Whole-Genome Level

We used a total of 4,391,835 SNPs detected from 254,721,076 aligned positions, after excluding SNP positions with missing data in any of the lines, to perform a global variability analysis. Global nucleotide diversity ($\pi_{tot} = 0.0066$) was among the highest reported in crop species (Qi et al. 2013). Within the melon varieties analyzed, the improved lines (elite), PS and Védrantais (VED), were the ones with the lowest diversity ($\pi_{total\_Elite} = 0.0035$), the cultivated landraces had an intermediate value ($\pi_{total\_Landrace} = 0.0052$), and the wild ecotypes the most diverse ($\pi_{total\_Wild} = 0.0094$), which is concordant with comparisons of cultivated and wild varieties in other species (Qi et al. 2013). Over the complete set of samples, the whole-genome synonymous diversity ($\pi_{syn} = 0.0055$) was lower than the total silent diversity ($\pi_{sil} = 0.0074$). Although the variability at silent positions is assumed to be constrained on regulatory regions and on unknown functional positions (Mackay et al. 2012), those are only a small fraction of the silent positions. In addition, synonymous positions can also be constrained due, for example, to codon usage preference (Ingvarsson 2010). Nonsynonymous diversity ($\pi_{nsyn} = 0.0015$) was lower (4-fold) than the diversity at synonymous positions, as expected. The nucleotide diversity was distributed nonuniformly across the genome, being generally lower at the extremes of the chromosomes (toward the telomeres) and higher in the pericentromeric regions (fig. 1A).

The wild accessions Cabo Verde (CV) and Trigonus (TRI) showed the highest number of unique SNPs and DIPs (table 1). A pairwise comparison between varieties showed a clear pattern of differentiation of the CV line versus the other six lines, CV having many more differences versus the rest of the lines that any other pairwise combination (supplementary table S3, Supplementary Material online). This suggests that CV may have been isolated from the rest during a long period. Principal component analysis also showed CV clearly separated from the rest in the first principal component, whereas TRI separated from the remaining five lines with the second principal component (supplementary fig. S2, Supplementary Material online).

Nucleotide Divergence in Melon Lines Compared with Cucumber and Melon Population Structure

Melon and cucumber lineages split 10.1 Ma (Sebastian et al. 2010), and cucumber is the closest relative to melon with an available genome sequence. For this reason, we used resequencing data from a breeding cucumber line, mapped to the melon reference, to analyze the divergence of melon versus cucumber (*Cucumis sativus* L.). This analysis was restricted to positions found within regions that can be aligned between the two species, totalling to 117,514,001 positions (46% of the total number of positions used for the variability analyses within the melon varieties), and detected 1,264,489 SNPs. The level of diversity per nucleotide for all the categories detected was relatively low ($\pi_{tot} = 0.0041$, $\pi_{sil} = 0.0049$, $\pi_{syn} = 0.0045$, $\pi_{nsyn} = 0.0010$), which was as expected as only conserved regions are included. The analysis showed a clear differentiation (the divergence per position corrected by Hasegawa–Kishino–Yano [HKY] is $K = 0.034$), indicating that it is suitable as outgroup, and can be used for polarizing melon ancestry from the derived variants.

The genealogical representation using the matrix of pairwise distances for the whole genome using cucumber as an outgroup (fig. 2A) shows that the ssp. *melo* forms a monophyletic group, whereas the *agrestis* group is a heterogeneous group, with Calcuta (CAL) closer to the *melo* clade and CV being the more divergent line. The topology shown is supported by a statistical analysis shown in supplementary figure S3, Supplementary Material online. A recent study of 93 melon accessions, including 75 *melo* and 18 *agrestis* types,

Table 1. SNPs and DIPs between the Seven Melon Lines and the DHL92 Reference Genome.

SNPs and DIPs	CV	IRK	PS	SC	TRI	CAL	VED	Total
SNPs	2,189,790	1,110,612	835,481	679,942	1,281,217	1,439,763	1,331,441	4,556,377
DIPs	208,431	132,123	98,959	81,756	165,699	173,250	145,460	718,832
Unique SNPs	842,870	59,513	41,365	93,199	221,946	85,345	68,306	
Unique DIPs	69,884	6,194	3,549	8,860	23,348	9,553	6,174	

NOTE.—Unique variants are those unique to that given line.

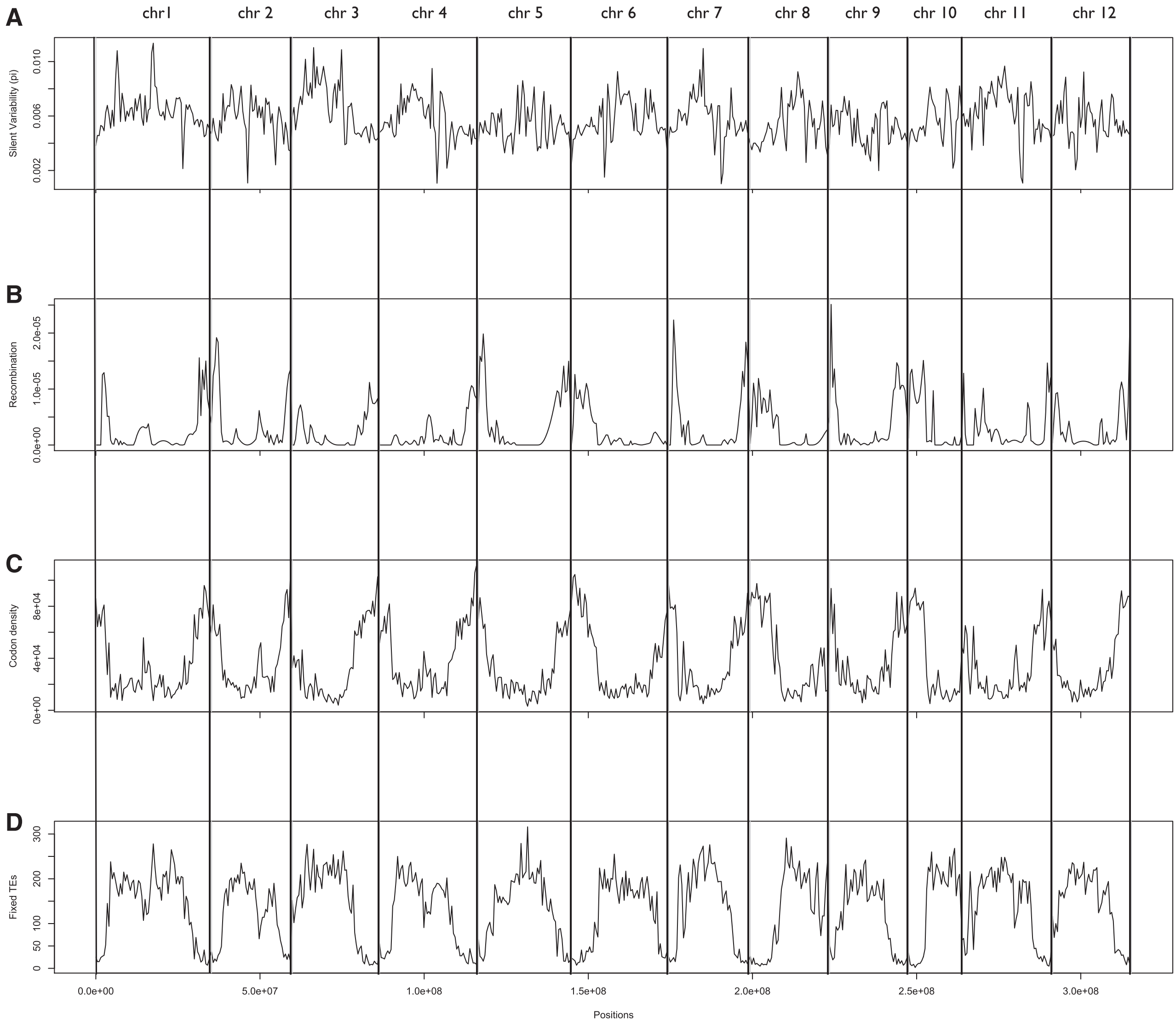


Fig. 1. (*A*) Nucleotide diversity of melon (π) across the genome in windows of 500 kb. (*B*) Distribution of the recombination rate across the genome in windows of 500 kb. (*C*) Distribution of codon density across the genome in windows of 500 kb. (*D*) Number of fixed TE across the genome in windows of 500 kb.

showed a strong population structure with different levels of admixture depending on the melon type, and identified two *melo* (*inodorus* and *cantalupensis*) and three *agrestis* (Indian *momordica*, African *agrestis*, and Far-eastern *conomon*) subpopulations (Esteras et al. 2013). Five of the seven melon accessions used here were included in this study, and CAL was also located closer to the *melo* accessions than the rest of the *agrestis* accessions. A per chromosome phylogenetic reconstruction (supplementary fig. S4, Supplementary Material online) showed relatively consistent nodes among all chromosomes, being PS-VED, PS-VED-IRK-CAL, and TRI-SC common groups across the genome. Note that the PS-VED-IRK node (ssp. *melon*), although common, is usually mixed with CAL lineages, suggesting a recent possible introgression or close recent ancestors. To analyze in more detail the general history of these lineages, we calculated neighbor-joining trees using windows of 10 and 100 kb across the whole genome. We observed 5,480 different topologies when using windows of 10 kb (1,185 in windows of 100 kb) across the whole genome and from a total of 30,812 (3,169) windows (supplementary fig. S4, Supplementary Material online). Up to 120 (117 in 100 kb bins) different nodes (clusters of lineages) were observed. Note that the maximum number of topologies for seven lineages using a rooted tree is 10,395 (Felsenstein 1978). Although the percentage of support of each node at each subset tree seems low (no more than 30% in the best case), the tree based on the whole genome is highly robust because its branches are the most frequently observed (Tree Rank Number, TRN = 3, see the definition of this indicator in Materials and Methods) and their bootstrap support high. The large number of observed window topologies (gene trees) is expected for lineages from the same species due to recombination during the history of the species. On the other hand, the number of topologies differs significantly from being obtained by chance (*P* value ≪ 1e-6), indicating a consistent population structure in the melon species.

The Role of Selection at Amino Acid Positions

Several plant species as *Helianthus annuus* (sunflower), *Populus tremula*, or *Capsella grandiflora* show evidence of positive selection at nonsynonymous positions (Fay 2011) whereas this was not the case in many others (e.g., crop species such as *Zea mays*, *Sorghum bicolor*, and *Oryza sativa*). In order to test whether the melon genome is evolving under selection we performed a MacDonald–Kreitman test (MKT, [McDonald and Kreitman 1991]) that tests whether the ratio between variability and divergence is different over synonymous and nonsynonymous positions by means of a Fisher exact test. We detected a very significant result of MKT (MKT = 36.3, *P* value = 1.7E-9), suggesting that, indeed, the melon genome as a whole has evolved under selection. We calculated the proportion of nonsynonymous positions affected by positive selection as $\alpha = 1-(K_{syn}\ \pi_{nsyn})/(\pi_{syn}\ K_{nsyn})$. The value of α was estimated as −0.24, indicating that negative selection predominates in the evolution at nonsynonymous positions (supplementary table S4, Supplementary Material online). α was very negative when calculated for only polymorphic singletons ($\alpha = -0.31$), indicating that many of the low frequency amino acid variants were negatively selected, but was also negative for the higher variant frequencies ($\alpha = -0.19$ for variants at the highest frequency, that is, derived mutation at six samples [Messer and Petrov 2013]). This indicates a global negative effect of selection on nonsynonymous positions throughout the history of melon.

The Role of Selection Considering the Patterns of Association of Genomic Features with Polymorphism and Divergence

We sought to determine whether the evolution of the melon genome has been mainly subjected to positive, neutral, or background selection. For this analysis, we assumed that genomic features such as recombination and gene density have structural patterns associated to the whole species and therefore, that there are not important differences in the gene density distribution neither in the recombination levels between different lineages nor between populations. Moreover, the distribution of the variability across the genome over different groups is rather similar, so we considered studying jointly the species sample. In melon, the recombination rate estimates (Argyris et al. 2015) and the codon density content show an accused nonuniform distribution, as they concentrate in the distal parts of the chromosomes (fig. 1*B* and *C*). We analyzed the distribution of silent (synonymous and noncoding) variability relative to coding region densities and to recombination rate and found that it is negatively associated with codon density (partial correlation $R = -0.37$, P value $\ll$ 1E-15) but not with the recombination rate (partial correlation $R = -0.02$, NS, fig. 3). This pattern does not fit with a strict neutral model of evolution, which predicts no association with these two features. This could be due to a reduction of variability in regions rich in coding regions (selectively constricted at nonsynonymous positions) or a mutational bias, where high numbers of mutations were located at low codon density regions. We did not detect association of variability with the percentage of GCs when considering codon density (supplementary fig. S5, Supplementary Material online), suggesting that a mutation bias for this dinucleotide is not responsible for the patterns of variability observed. On the other hand, we did not observe a negative association of synonymous polymorphisms with nonsynonymous divergence that would be expected for evolution under positive selection (recurrent selective sweeps) (fig. 4). Our data suggest an evolution under background selection. However, the expected positive correlation of neutral (silent) variability with recombination under this model is not observed (fig. 3*B*). Negative or no association (as it is our case) of neutral variability with recombination rate has already been described in other plant genomes and has been explained by a nonuniform distribution of coding regions that would generate a nonuniform distribution of selective mutations (Flowers et al. 2012).

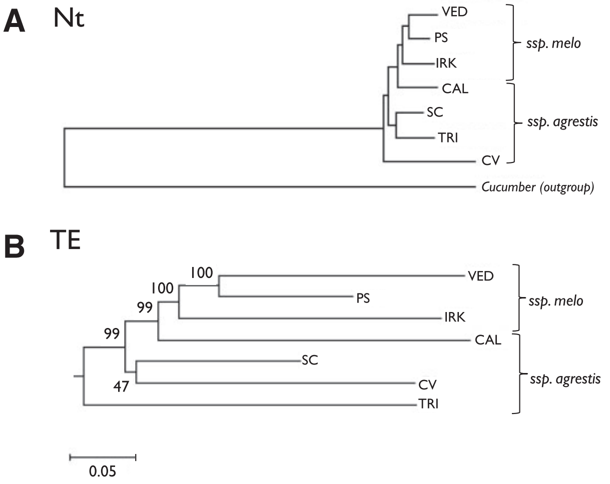


Fig. 2. SNP-based (*A*) and transposon-based (*B*) phylogenetic relationships among the seven resequenced lines. Bootstrap support values are shown in black numbers.

The distribution of nucleotide divergence along the genome is nonuniform, being lower in pericentromeric regions and higher towards the telomeres (supplementary fig. S6, Supplementary Material online). The neutral model also predicts a positive association between variability and divergence. Our results show that there is a negative association of silent polymorphism with divergence (Kendall tau = −0.33, supplementary fig. S7, Supplementary Material online, see also fig. 1*A* and supplementary fig. S6*A*, Supplementary Material online). This association is also negative and very significant when considering codon density and recombination (supplementary fig. S8*A*, Supplementary Material online), and in contrast to the patterns of polymorphism versus divergence at synonymous (Kendall tau = +0.13) and nonsynonymous (Kendall tau = +0.26) positions (supplementary fig. S8*B* and *C*, Supplementary Material online), suggesting that silent positions may not behave as neutral. This pattern is unusual, in rice a positive or null association between polymorphism and divergence was found (Flowers et al. 2012). On

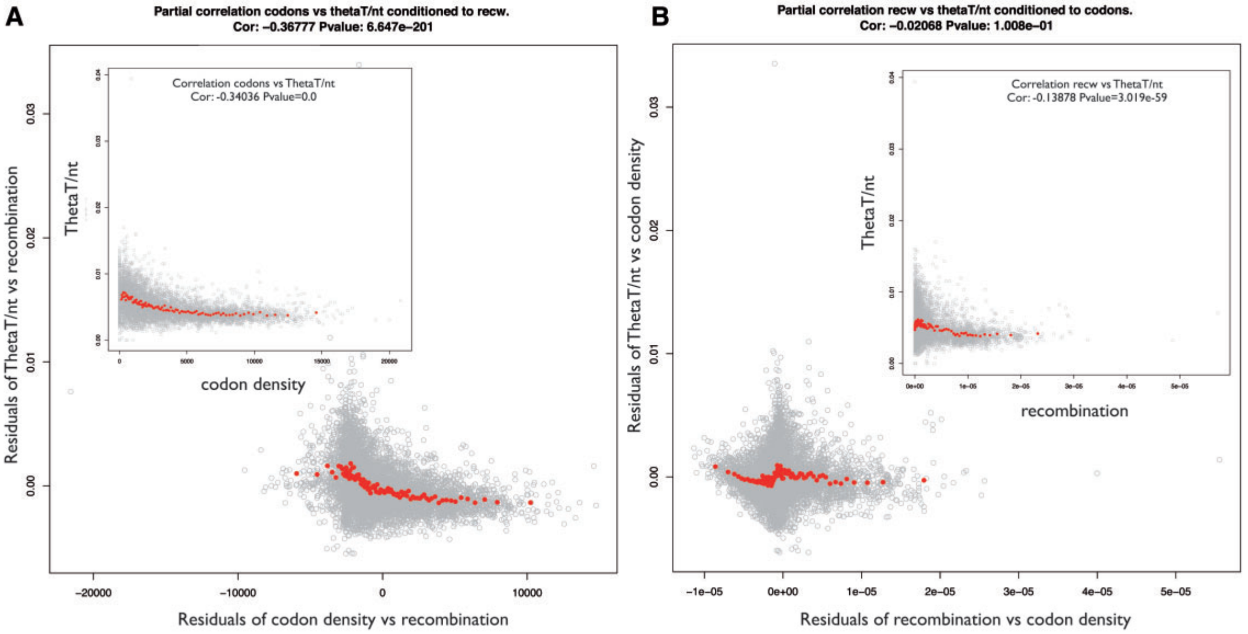


FIG. 3. The left panel (*A*) shows in the inner plot the correlation of silent polymorphisms versus codon density and in the external plot the same correlation but considering recombination in the partial correlation (R = -0.37, P value = 6.4E-201). The right panel (*B*) shows in the inner plot the correlation of silent polymorphisms versus recombination and in the external plot the same correlation but considering codon density in the partial correlation (R = −0.02, P value = 0.1).

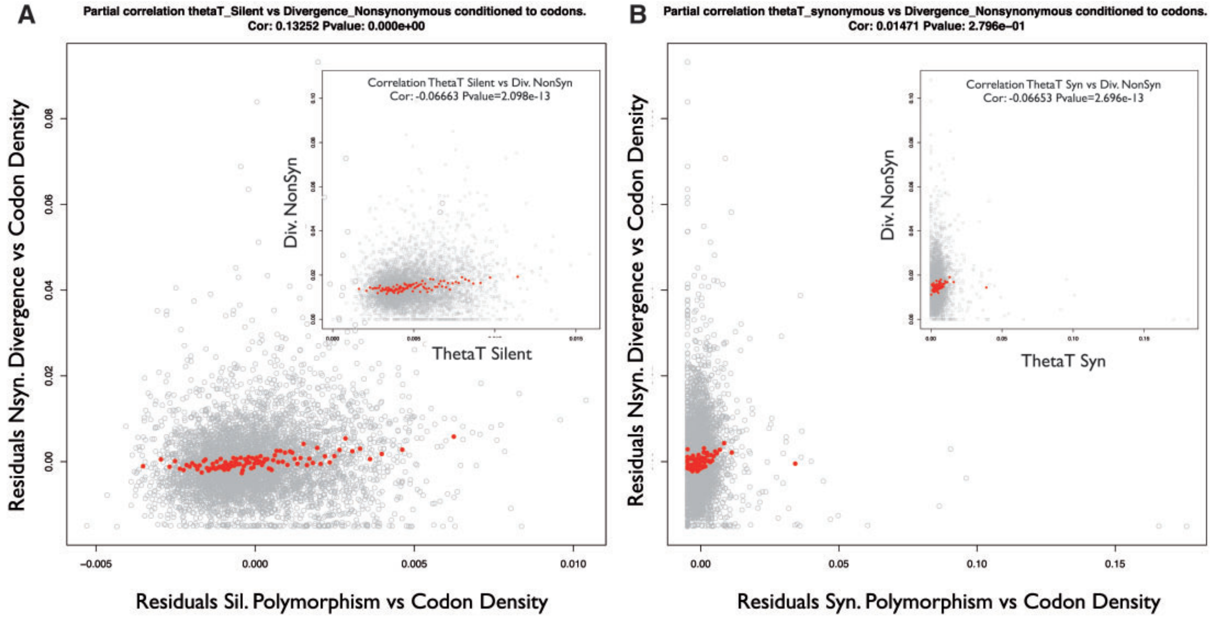


FIG. 4. The left panel (*A*) shows in the inner plot the correlation of silent polymorphisms versus nonsynonynmous divergence and in the external plot the same correlation but considering codon density in the partial correlation (R = 0.13, P value < 1E-200). The right panel (*B*) shows in the inner plot the correlation of synonymous polymorphisms versus nonsynonymous divergence and in the external plot the same correlation but considering codon density in the partial correlation (R = 0.01, P value = 0.28).

the other hand, in regions of high coding density or near genes we observed no association of polymorphism with divergence (supplementary fig. S9, Supplementary Material online), which is similar to what has been previously reported in rice (Flowers et al. 2012), where they used silent regions close to genes. The intriguing silent divergence pattern observed may be due (total or partially) to the difficulty to align regions with low gene density. The ratio of unaligned positions with the outgroup is highest in the pericentromeric regions (~80%) and lowest at rest of chromosomal arms (~40%). Although this difference can be explained by the rapid expansion of TE elements across the melon genome after speciation (mostly at pericentromeric regions, see the following sections), only the 62 % of the cucumber genome was successfully aligned with melon, indicating that only the vicinity of the highest conserved regions were taken into account for this comparison.

Regions of Interest for Their Singular Patterns of Variability and Differentiation

As a first step toward identifying the regions of the melon genome that have been selected during its recent evolution, we analyzed the patterns of variability and divergence between different sets of varieties, including elite (VED and PS) versus nonelite (SC, TRI, CAL, IRK, and CV) and *melo* (PS, VED, and IRK) versus *agrestis* (CAL, SC, TRI, and CV) subspecies.

We observed high divergence across the whole genome when comparing *melo* versus *agrestis* subspecies. High diversity at the *melo* group was located at chromosome 7 but also coincident with a peak in *agrestis* (supplementary fig. S10, Supplementary Material online). The comparison of elite (VED and PS) versus nonelite varieties (SC, TRI, CAL, Irak-IRK-, and CV) shows that the total variability (π) is extremely reduced within the elite group in chromosomes 1 and 6 (supplementary fig. S11, Supplementary Material online) but very high at others (chromosomes 3 and 8). Of particular interest is a region in chromosome 1 showing almost no variability among the elite varieties while being high among the nonelite varieties (supplementary fig. S11, Supplementary Material online).

An analysis of the fixation index *F*st (Hudson et al. 1992) per chromosome showed clear differences when comparing the varieties belonging to the *melo* and *agrestis* subspecies, and also when comparing the elite highly inbred lines to the rest (nonelite) (supplementary fig. S12 and table S5, Supplementary Material online). The highest difference was found between the elite and nonelite varieties (mean *F*st = 0.191). These two elite lines (PS and VED) are the closest lines among those analyzed and therefore have the lowest variability (supplementary fig. S11*A*, Supplementary Material online) ($\pi_{\text{Elite}} = 0.0035$, $\pi_{\text{NO-Elite}} = 0.0070$). The comparison of the *melo* and *agrestis* groups also showed that the former is more homogeneous ($\pi_{\text{melo}} = 0.0041$, $\pi_{\text{agrestis}} = 0.0076$).

In order to look for particular regions of differentiation, we also analyzed the variability in 500 kb windows across the genome. The level of variability depends on the chromosome location, being generally higher at pericentromeric regions. In contrast, *F*st is midly associated to codon density and it is not associated to recombination (supplementary fig. S13, Supplementary Material online), although heteroscedasticity is observed (that is, heterogeneity in the variance across the codon density parameter). In order to choose the more significant values, we used a variable threshold value in relation to codon density (traced assuming two times the standard deviation calculated in bins of ten fragments, which is 99% in a normal distribution), because their variance is different across this variable (supplementary fig. S14, Supplementary Material online). *F*st values above the threshold curve may be associated to regions involved in the differentiation between these two groups. In the comparisons of *melo* versus *agrestis* subspecies, the more extreme population differentiation regions (supplementary fig. S14*B*, Supplementary Material online) were located at chromosomes 1 (positions 12e6 to 12.5e6, 19e6 to 20e6, 28.5e6 to 29.5e6), 3 (positions 21e6 to 21.5e6), 6 (positions 13e6 to 13.5e6 and from 17.5e6 to 18e6), 7 (positions 13e6 to 13.5e6), 8 (positions 1 to 0.5e6), and 11 (positions 11e6 to 11.5e6 and 21.5e6 to 22e6). For the comparison among elite and nonelite varieties, the outlier windows for *F*st-index considering codon density are located at chromosome 1 (positions 3e6 to 3.5e6), 2 (positions 24e6 to 24.5e6), 3 (positions 21e6 to 21.5e6), and 6 (positions 25e6 to 26e6).

Nontransposon Structural Variation

The most common methods used to discover SVs are based either on discordant mapping signatures of paired reads or by variations in read-depth (Alkan et al. 2011). We developed an in silico workflow that implements these two approaches to identify SV in the seven melon lines with respect to the reference, and combined, have identified 3,609 SVs. SVs were classified as deletions or PAV ($n = 2{,}541$), inversions ($n = 620$), and duplications or CNV ($n = 448$) (supplementary table S6, Supplementary Material online). The number and types of SV found are in general consistent with what has been reported in other plant species (Saxena et al. 2014), although the use of different methods and parameters in different studies make them difficult to compare. A total of 902 genes are affected by a SV in at least one variety (supplementary table S7, Supplementary Material online). Of these, 745 fell in deletions, 142 in tandem duplications, and 15 in inverted regions. According to the public available melon gene ontology (GO) functional annotation (Garcia-Mas et al. 2012), refined in this paper using annotation with automated assignment of human readable description (AHRD), 53 genes were related to agronomically relevant pathways, including disease resistance (29), cell-wall metabolism (10), aroma volatiles metabolism (9), sugar metabolism (4), and carotenoid biosynthesis (1).

The five largest deletions were found to range between 82 and 416 kb long (supplementary table S8, Supplementary Material online). One of these large deletions is located in chromosome 5 and spans 146 kb in CV and 82 kb in SC, affecting six resistance genes (R-genes) of the NBS-LRR class (MELO3C04318-4324). This deletion is found in a 1.1 Mb region of chromosome 5 where the largest R-gene cluster in the melon genome is located, containing 23 R-genes (Garcia-Mas et al. 2012; González et al. 2014). The same deletion in CV and SC was also described in a recent presence/absence gene variability study using a subset of our melon lines (CV, SC, PS, and IRK) and a different discovery pipeline (González et al. 2013). Additional R-gene clusters are affected by SV, namely the MELO3C004289-4295 interval in the vicinity of the above-mentioned deletion in chromosome 5 (supplementary table S7, Supplementary Material online, in yellow) and another in chromosome 1 (MELO3C023566-23578) (supplementary table S7, Supplementary Material online, in yellow). A similarly high variability in resistance gene clusters has been reported recently in soybean using array hybridization and targeted resequencing (McHale et al. 2012). Four out of the five large deletions described in supplementary table S8, Supplementary Material online, were also detected in González et al (2013), confirming the accuracy of our SV discovery pipeline.

Transposon Insertion Polymorphisms

Mobile elements are an important source of the variability necessary for evolution (Lisch 2013; Olsen and Wendel 2013). In order to analyze the contribution of transposon insertions to the genotypic variability in melon, we first refined the transposon annotation we previously performed (Garcia-Mas et al. 2012) with an additional search for miniature inverted terminal-repeat elements (MITEs) using Subotir (Henaff et al. 2014) and MITE-Hunter (Han and Wessler 2010). MITEs are a particular type of transposons abundant and active in plant genomes and therefore it was important to include them in the annotation (Casacuberta and Santiago 2003). The previously performed annotation included MITEs related to other annotated class II elements (these are included within the different class II TEs superfamilies, see table 2); with this additional search the annotation also includes MITEs nonrelated to the previously detected families of class II elements (these MITEs are classified as "other MITEs," see table 2). We used a combination of publicly available programs and tools newly developed in our laboratory to identify transposon deletions and insertions in the resequenced melon varieties with respect to the reference genome. Tools are routinely compared with others in the context of methods papers, sometimes with simulated data (Layer et al. 2014), or real data sets that include gold standard variants from projects such as the 1000GP or GIAB (Rausch et al. 2012) but the performance of these depends greatly on the sequencing depth and complexity of the reference. To date there is no reference data set of gold standard variants in plants, and the 1000GP data used to benchmark the gold standard calls is much lower coverage and shorter reads than our data set. In order to evaluate the performance of the programs to be used on our data set, we took advantage of the unique possibility offered by the fact that we have resequencing data for the same line that was used to generate the reference sequence. We generated a simulated reference genome by deleting 1,871 transposons from the assembled DHL92 reference genome and inserting them in randomly chosen locations (see supplementary material S1, Supplementary Material online). These deletions and insertions can be used for benchmarking as they should be detected as insertions and deletions, respectively, in the same DHL92 line mapped to the modified reference, and is the most accurate measure of sensitivity and specificity as we model the complexity of the reference genome and the characteristic of our sequencing data sets. We evaluated BreakDancer (Chen et al. 2009) and Pindel (Ye et al. 2009) in their ability to detect deletions with respect to the reference. In our hands (see Materials and Methods section for details), Breakdancer has a sensitivity of 79 % and a positive predictive value (PPV) of less than 64 %. Pindel shows comparable sensitivity (76 %) yet higher PPV (85 %) (supplementary table S9, Supplementary Material online). Therefore we chose to use Pindel to detect TE deletions with respect to the reference genome.

To detect insertions in the resequenced genomes with respect to the reference, we used a program recently developed in our laboratory which relies on discordant and soft-clipped read signatures to predict TE insertion loci, named Jitterbug (Hénaff et al, submitted). Using the previously described simulation, we turned the parameters to achieve 82.71% sensitivity and 98.68% PPV with our data set. We then ran Jitterbug on the data for all seven lines, and identified a total of 2,688 insertions, consisting in 2,056 polymorphic loci (corresponding to an insertion at the same locus in one or more lines). In order to confirm this high sensitivity and specificity values, we analyzed by polymerase chain reaction (PCR) 23 of the predicted polymorphic loci amongst the seven lines, with primer sets designed to detect both the TE insertion and the reference allele (or empty locus). All the 23 polymorphic loci were confirmed by PCR (supplementary fig. S15, Supplementary Material online). However, while in 20 cases the genotype predicted for the seven varieties was confirmed by PCR, in one case we detected the empty site for one of the varieties that was supposed to contain the insertion and in three additional cases we failed to amplify the region in some varieties that were predicted to not contain the insertion (supplementary fig. S15, Supplementary Material online). These discrepancies may be due to a lack of fixation of the insertion within the population, as the DNA used for

Table 2. Breakdown of TE Families for TE-Related Polymorphic Loci (PM).

Superfamily	Number of PM	Percentage of PM	Number of Copies in the Genome	Percentage of Copies in the Genome
Gypsy	661	24.17	28,174	23.68
Copia	635	23.22	17,346	14.58
Non-LTR retrotransposon	10	0.37	129	0.11
Retrotransposon fragment	469	17.15	42,733	35.91
Total Retrotransposons	1,775	64.90	88,382	74.27
CACTA[a]	93	3.40	6,944	5.84
hAT[a]	11	0.40	677	0.57
MULE[a]	239	8.74	9,497	7.98
Mariner[a]	3	0.11	609	0.51
Other MITEs	355	12.98	2,175	1.83
Helitron	4	0.15	746	0.63
PIF[a]	147	5.37	3,094	2.60
DNA TE fragment	29	1.06	6,874	5.78
Total DNA TE	881	32.21	30,616	25.73
Uncategorized	79	2.89		
Total	2,735	100.00	118,998	100.00

[a]Including short elements that could be MITEs.

sequencing and PCR analysis came from individuals different from those used for resequencing experiments. Indeed, although the elite PS and VED varieties are highly inbred and homozygous, the wild varieties and landraces may show a higher degree of heterozygosity and TE insertions may segregate in the population. A clear example is CAL which is heterozygous for the insertion in at least four of the polymorphic sites surveyed by PCR. The lack of amplification in PS of the empty site corresponding to the CM_4552 locus is probably the result of an insertion in PS that was not predicted by Jitterbug due to a low resequencing coverage in this variety at this particular location (not shown).

Using Pindel and Jitterbug to call deletions and insertions, respectively, in the resequenced varieties with respect to the reference we detected 2,735 polymorphic TE insertions. Two-thirds of these polymorphisms were caused by retrotransposons, DNA transposons insertion/deletions roughly accounting for the remaining one-third (we were not able to categorize 3.3% of the polymorphic sites due to their complex nature) (table 2). Among retrotransposon-related polymorphisms, most of them were contributed by *copia* (23%) and *gypsy* (24%) elements, whereas most of the DNA transposon-related polymorphisms were contributed by MITEs (table 2). A small number of TE families were responsible for the large part of the observed polymorphisms. Indeed, nine families, which represent less than 4% of the TE copies annotated in the melon genome, account for more than one-third of the polymorphic TE insertions (table 3). Judging from their sequence similarity, these families contain relatively young elements (not shown), which is consistent with their recent activity during melon domestication and breeding. It is interesting to note that as much as 60% of the polymorphic TEs are present in only one variety, which is also consistent with a recent TE activity (supplementary table S10, Supplementary Material online). We used the TE insertions shared by more than one variety to construct a dendrogram of the phylogenetic relationships of the seven melon varieties. We used a NJ approach to obtain a dendrogram as the nonconstant evolutionary rate of TEs is not consistent with the UPGMA approach. Indeed, it is generally accepted that the activity of TEs, and in particular that of long terminal repeat-retrotransposons and MITEs, is not constant over time, with burst of transposition being followed by periods of relatively low activity (Lu et al. 2012; El Baidouri and Panaud 2013). The dendrogram obtained with the TE data shows a pattern consistent with the dendrogram based on SNPs, although the relative length of the branches are very different (much longer at external nodes), possibly by the faster and nonconstant rate of evolution of TEs (fig. 2*B*).

Table 3. Most of the Polymorphisms Were Caused by the Mobilization of a Small Number of Transposon Families.

Family	Superfamily	PM Sites	%	Annotated	%
CM_MITE_2617	CACTA (MITE)	224	8.19	700	0.59
CM_MULE_10	MULE	187	6.84	682	0.57
CM_gypsy_116	Gypsy	120	4.39	177	0.15
CM_PIF_6	PIF	111	4.06	881	0.74
MELON_MITEs_ 1_43749	PIF (MITE)	110	4.02	117	0.1
CM_copia_96[a]	Copia	70	2.56	1,695	1.42
CM_copia_45	Copia	59	2.16	184	0.15
CM_copia_70[a]	Copia	59	2.16	39	0.03
CM_gypsy_137	Gypsy	51	1.86	107	0.1
Total		991	36.23		3.85

[a] Complex families composed of nested insertions.

The annotation and analysis of the transposons present in the melon reference genome, and the comparison of these data with the transposon content in cucumber, showed that transposons have been very active during melon recent evolution (Garcia-Mas et al. 2012), transposing and amplifying to a greater extent in melon compared with cucumber. The results presented here confirm the recent transposon activity in melon genome and suggest that transposons may be at the origin of an important fraction of the variability in this species.

Transposon density usually shows a nonrandom distribution along plant chromosomes, with TEs concentrating in the pericentromeric regions where gene density is lower (*Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Paterson et al. 2009; Schnable et al. 2009; Schmutz et al. 2010; Tian et al. 2012; Tomato Genome Consortium 2012; Choulet et al. 2014). In plants these regions frequently show a low recombination rate and it has been proposed that this may also explain the higher concentration of TEs in these regions as they may be more difficult to eliminate by selection (Gaut et al. 2007). The distribution of fixed (annotated in the reference and nonpolymorphic) TEs in melon is nonrandom, with higher density in large regions flanking the centromeres (Garcia-Mas et al. 2012) (fig. 1*D*). We confirmed that fixed TEs show an inversely correlated distribution with respect to gene density as well as to recombination rate, and we tested which of these two features influences TE distribution. Our results show that the density of fixed TEs is strongly negatively associated with codon density (as seen in partial correlations, supplementary fig. S16, Supplementary Material online) indicating an accumulation of fixed TEs in nonfunctional regions. On the contrary, the frequency of fixed TEs shows no association with recombination rate (supplementary fig. S16, Supplementary Material online).

The frequency of polymorphic TEs across the genome is more homogeneously distributed (supplementary figs. S17 and S18, Supplementary Material online) than that of SNPs. However, when using the theta estimate of Zeng et al. (2006), which weights more the high-frequency variants, it can be seen that there is a slight bias to pericentromeric regions (supplementary figs. S17 and S18, Supplementary Material online). This pattern suggests the effect of selection on TE distribution, possibly eliminating those elements that affect functional regions and allowing their fixation in regions with less functional constraints. Alternatively, this could also be the consequence of an insertion preference within pericentromeric regions of some TE families. Indeed, it has been previously shown that some TE families, in particular among retrotransposons, target pericentromeric regions for insertion and accumulate almost exclusively within these regions (Peterson-Burch et al. 2004; Du et al. 2010; Sharma and Presting 2014).

We found generally lower TE diversity in elite varieties when compared with the rest (supplementary fig. S11*B*, Supplementary Material online). Indeed, there are 613 polymorphic TE insertions between VED and PS, whereas there are 2,092 polymorphic sites among nonelite varieties, and there is no combination of varieties showing a level of polymorphisms comparable or smaller than that of the two elite varieties ($p = 0$). This low TE diversity is particularly clear in some chromosomes and chromosomal regions such as portions of chromosomes 1 and 6. For example, an 11 Mb region (positions 5e6 to 16e6) in chromosome 1 shows no TE polymorphisms between elite varieties and 99 TE between the nonelite. This number is significantly lower ($p = 0$) than any other region for the elite varieties, and not significantly different ($p = 0.99$) for the nonelite varieties. Interestingly, these regions also show very low nucleotide diversity (supplementary fig. S11*A*, Supplementary Material online) which suggests that these regions may have been fixed during breeding. An analysis of the 306 genes found in this region that have an associated GO term (of the 532 genes present in this region) shows enrichment in GO terms related to cellulose biosynthesis (P value $= 2.2 \times 10^{-5}$). Although cellulose synthesis and breakdown may be related with fruit softening and this is an important trait for melon breeding, more work is needed to evaluate the biological significance of this finding.

The analysis of the *F*st index (supplementary fig. S19, Supplementary Material online) measuring differences between elite and nonelite showed several peaks of high differentiation, also suggesting that some TE insertions may have been fixed during the breeding process of these two elite lines.

In general, there is an important fraction of the polymorphic TE insertions located in genic regions, suggesting a potential impact on genes. Indeed, more than 22% of the 2,735 polymorphic TEs are located within an annotated gene, and an additional 7.8% are located within 500 nt of a gene. Among TEs located within genes, 361 are within exons and 250 in introns and untranslated regions (table 4). This data set should allow in the future to evaluate the impact of transposons on the evolution of the melon genome. Of particular interest is the fact that we identified 165 TE insertions in coding regions that are polymorphic between the two elite lines VED and PS. These two elite lines are closely related phylogenetically (fig. 2*A*), but still they show important differences in key agronomical traits such as fruit shape, flesh color, ripening behavior, sugar content, and aromas. The possibility that one of these TE insertions within genes may have altered one of these characteristics is highly likely. In fact, an analysis of the genes showing transposon insertion polymorphisms shows that an important fraction of them are related to the development of reproductive structures, hormone signaling, and sugar metabolism, suggesting that transposon insertions may have modified some of the metabolic pathways or regulatory networks that underlie these important agronomic traits.

Conclusions

We present here a pioneer work in plants consisting of a comprehensive analysis of variability in a crop species, from SNPs to large SV, including transposons insertion polymorphisms, and which includes new tools and bioinformatic pipelines to integrate these analyses. Our benchmarking of these algorithms using the resequence of the reference genome is a novel approach and ensures an accurate estimation of the specificity and sensitivity of the results for our specific data set. In particular, our assessment of Jitterbug shows that it is a very specific new tool for TE insertion polymorphisms identification.

The variability found among seven melon varieties that represent the extant diversity of the species and include wild accessions and breeding lines, is relatively high due in part to the structure of the species. The nucleotide diversity is distributed nonuniformly across the genome, being generally lower at the extremes of the chromosomes, coinciding with gene-rich and high-recombination regions, and higher in the pericentromeric regions, where gene density and recombination rate are low and there is a higher accumulation of TEs. However, this variability is greatly reduced among elite varieties, probably due to the selection during breeding. We have found some chromosomal regions that show a high differentiation of the elite varieties versus the rest of the varieties analyzed, which could be considered as regions that suffered strong selection. Interestingly, some of these regions also show a high differentiation between elite and nonelite varieties with respect to polymorphic TE insertions suggesting that these regions may have been fixed during breeding. Our data also suggest that transposons may be at the origin of an important fraction of the variability in melon, in addition to the variation due to SNPs and SVs. As much as 60% of the polymorphic TEs are present in only one variety, suggesting that there have been an important transposon activity very recently during melon evolution.

Additionally, a total of 902 genes were shown to be affected by a SV in at least one variety, with a significant enrichment in regions harboring R-gene clusters. We found that the largest R-gene cluster in the melon genome, located in chromosome 5 and comprising 23 genes, has been partially deleted in some of the accessions.

As resequencing costs decrease, the analysis of large data sets of varieties that represent the extant variability of crop species is becoming feasible. However, up to now these analyses have been restricted to SNPs and, in few cases to large SV. The approach presented here describes a comprehensive analysis of variability including SNPs, SVs, and also TE insertion polymorphisms, and implements the in-depth variability analysis that can be used to detect genomic regions involved in domestication and selection when the resequence of a wide collection of melon germplasm is available.

Materials and Methods

Plant Material

Seven melon accessions were used in this study, three from the ssp. *melo* and four from the ssp. *agrestis* (supplementary

Table 4. TE Insertions Located in Genic Regions.

	Total PM Sites	PM Sites < 500 bp	%	PM Sites in Genes	%	PM Sites in Exons	%
All lines	2,735	826	30.20	611	22.34	361	13.20
agrestis versus *melo*	31	4	12.90	4	12.90	3	9.68
Elite versus nonelite	69	13	18.84	12	17.39	8	11.59
PS versus VED	671	231	34.43	165	24.59	105	15.65

table S1, Supplementary Material online). The ssp. *melo* accessions were the Piel de sapo line T111 (PS) (*inodorus* group), the cantaloupe type Védrantais (VED) (*cantaloupensis* group), and the C-1012 cultivar (IRK) (*dudaim* group). The ssp. *agrestis* accessions were PI 161375 (SC) (*conomon* group), PI 124112 (CAL) (*momordica* group) and the accessions Ames-24297, previously classified as *Cucumis trigonus* (TRI), and C-386 from CV. According to morphological and agronomic data, CV/TRI, SC/CAL/IRK, and PS/VED may be considered wild, landrace, and elite lines, respectively. DHL92, a doubled-haploid line obtained from the cross between PS and SC, and which was used to obtain the reference genome sequence of melon (Garcia-Mas et al. 2012) was also included in the analysis as a control. Seeds from the eight accessions were planted in trays and plants were grown under the same greenhouse conditions as previously described (Eduardo et al. 2005). For library construction, young leaf samples were used for DNA extraction (Garcia-Mas et al. 2001), mixing leaves of five plants per accession except for CV and IRK, where a single plant was used. DNA for PCR analysis was extracted from different individuals than the ones used for library preparation.

Resequence Analysis and SNP Calling

Paired-end libraries with an average insert size of 500 bp were produced and sequenced with the Illumina Genome Analyser IIx technology at the Centre Nacional d'Anàlisi Genòmica (CNAG, Barcelona) (supplementary table S2, Supplementary Material online). On average, more than 30 million paired reads were obtained for each melon accession with a read length of 150 bp. Resequencing data of DHL92, PS, SC, IRK, and CV has been already described in previous works (Garcia-Mas et al. 2012; González et al. 2013).

An in-house pipeline called SUPERW (simply unified pair-end read workflow) was developed to create a dynamic and fast tool to analyze the variation data produced from the resequencing experiments (supplementary material S1 and fig. S1, Supplementary Material online). The SUPER pipeline and the filtering script SUPERRA were developed and used with the melon resequencing data. The melon reference genome used was v3.5 (melon_genome_pseudomolecules_V3.5), available at http://www.melonomics.net (last accessed July 22, 2015). The parameters used were 1) for the filtering and mapping step a read PRHED quality > 25, a minimum length of 35 bp, removing all the Illumina adaptors and a mapping quality of PHRED > 10, 2) for the variation calling step (SNPs and DIPs) a genotype quality ≥20, a locus quality > 30 and a minimum depth coverage of five reads for both small and large variations, 3) only DIPs up to 200 bp were kept, and 4) at each variable locus, the allele frequency (AF) of the variant was calculated as the ratio of reads supporting a homozygous or heterozygous state. Variants were filtered to keep those with an AF > 0.75. Several benchmarks were used to reach an optimal quality for the data produced. The resequenced line DHL92, the same variety used to assemble the melon reference genome, was used as control to remove all the variants caused by 454 sequencing errors (supplementary table S11, Supplementary Material online). The resequence of DHL92 enabled us to determine the false discovery rate of called SNPs as 2.05 × 10E-5 per bp. The same approach was used to calculate the false discovery rate between the reference genome and one of its parental lines, in a region inherited from SC in chr12, being 2.66 × 10E-5. After the quality filtering, only homozygous sites were considered.

Calling Large Structural Variants

Both pair-end and depth of coverage approaches were used to predict large SVs. The paired-end approach was used to calculate the SV from 200 bp up to 25 kb while depth of coverage was used to identify SV larger than 25 kb. The results of the two algorithms were merged in order to create a nonredundant set of large SVs. The pair-read approach of Pindel v2.4 (Ye et al. 2009) was used to call variations up to 25 kb. Alignment files created with bwa sampe and bwa aln (bwa v 0.7.0) without removing multiple mapped reads were used to extract deletions, duplications, small insertions, and inversions considering the difference between the pair-end distance and mapped distance. All SVs were filtered for a depth of coverage to require at least 5×. Moreover specific filters were added to each SV: 1) should have a PHRED quality of 20, 2) inversions should be longer than 200 bp, 3) deletions should have both forward and reverse supporting reads, and 4) SVs of different types that fell in same genomic region (conflict SVs) were removed. In addition, for each SV the genes that have suffered a variation were extracted. All genes in regions affected by SVs were tallied and functionally annotated using a tool to assign automated assignment of human readable descriptions, (AHRD v2, https://github.com/groupschoof/AHRD, last accessed July 22, 2015). AHRD is able to select descriptions that are concise and informative, using BLAST hits taken from searches against Uniprot/trEMBL, Uniprot/Swissprot, and TAIR10.

Population Genetic Analysis

Estimates of nucleotide variability were calculated using Achaz equations (Achaz 2009) for Watterson, Tajima's, Fu and Li, Fay and Wu, and Zeng's theta with the folded and unfolded frequency spectrum, if necessary. Patterns of variability were inferred from calculation of the following neutrality tests: Tajima's *D* (Tajima 1989), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000; Zeng et al. 2006). Population differentiation—Fst (Hudson et al. 1992)—was calculated between chosen groups. An Illumina Genome Analyser IIx genome resequence of a cucumber inbreed line, kindly obtained from Semillas Fitó SA, was used for the divergence studies. Fixed variants and nucleotide divergence was calculated between *C. melo* and *C. sativus* using the number of differences from the total positions. Divergence was corrected for multiple mutations using the HKY model (Hasegawa et al. 1985). *mstatspop* was used to calculate all these statistics (made available by the authors at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html, last accessed July 22, 2015).

The groups considered in this study were *C. melo*, ssp. *melo* (PS, VED, IRK), *C. melo* ssp. *agrestis* (CAL, SC, TRI, CV), "elite" (PS and VED), "nonelite" (TRI, CV, IRK, CAL, SC), "Landrace" (IRK, CAL, SC), "Domesticated" (PS, VED, IRK, CAL, SC), and "Wild" (TRI, CV).

The analyses were performed on total, silent, synonymous and nonsynonymous positions using the GTF annotation file, considering the whole genome but also calculating separately the statistics by windows of 50, 100, and 500 kb. In order to analyse a possible influence of read depth on the measured variability and differentiation, the correlation of the mean read depth in windows of 50 and 500 kb and the levels of variability and differentiation among populations was calculated. No correlation was found between read depth and differentiation. A very low correlation (−0.104, *P* value = 4.57e-35) was observed between read depth and variability, and no differences were observed when introducing read depth as a dependent factor in our analyses.

A table with the presence/absence of all annotated transposon elements (TE) was used to calculate the number and the frequency of these elements on the whole and on each desired window. Estimates of diversity (Watterson, Tajima, Fu and Li, and Zeng) were calculated with folded and with unfolded frequency spectrum for each window of size 50, 100, and 500 kb. Smaller windows showed higher variance in the studied statistics and were not used in global analysis. The frequency spectra and the Tajima's *D* test were also calculated to study the patterns of TE variability. Finally, population differentiation was also calculated between different groups or populations.

Kendall rank association values and their probabilities were calculated with the R-environment (http://www.rproject.org, last accessed July 22, 2015) to estimate the association between any two variables. Similarly to Cai et al. (2009), we calculated the mean of the variable located on the *y* axis on 100 separated bins for the variable located on the *x* axis. These values were plotted in red together with the whole values.

Partial correlation analyses were calculated assuming a normal distribution and comparing the residuals of the variables in relation to a third variable to account. Correlation and signification was obtained using the Pearson method.

The methods used for constructing dendrograms, performing Principal Component Analysis, estimating the proportion of nonsynonymous positions under selection and the analysis using recombination estimates are detailed at supplementary material S1, Supplementary Material online.

TE Insertion Polymorphism Analysis

Quality of reads was assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed July 22, 2015). Reads were filtered using SGA (https://github.com/jts/sga, last accessed July 22, 2015) with preprocess (-q 10 -m 50 –permute-ambiguous –pe-mode=1), then index (-d 2000000) and correct with default parameters. Reads were corrected and trimmed using SGA in order to maintain read pairs intact (as opposed to filtering performed for SNP analysis). Filtered reads were mapped to the assembled reference genome DHL92 v3.5 (Garcia-Mas et al. 2012) using BWA v 0.7.0 (Li and Durbin 2009) with *aln* (-n 6 -o 1 -e 1), and *sampe* with –s default parameter. SAMtools v 0.1.19 was used (Li et al. 2009) to sort and index all *bam* files.

Pindel v2.4 (Ye et al. 2009) and Breakdancer v1.2.6 (Chen et al. 2009) were used to detect deletions in the melon lines with respect to the reference genome. Pindel was run with a maximum range index of 5 and an anchor quality of 35. In order to decrease computational time, reporting of both inversions and duplications were disabled. Breakdancer was used with default parameters. Both sets of predictions were merged to remove redundancies, and an additional filtering step was applied to select those greater than 200 bp and less than 25 kb. Of these, those overlapping annotated transposons were retained for further analysis.

Analysis to detect transposon insertions was performed with Jitterbug (Hánaff E, Zapata L, Casacuberta JM, Ossowski S, submitted), using a value of 35 for the minimal mapping quality of the reads (-q 35). Jitterbug was parallelized using a bin size of 1 million, and insertions were filtered using the companion scripts supplied and the default parameters calculated on the fly by Jitterbug. The sensitivity and specificity of Pindel, Breakdancer, and Jitterbug was assessed on a simulated data set where a subset of annotated TEs we shuffled to random positions (see supplementary material S1, Supplementary Material online).

A subset of the predicted TE insertion polymorphisms were analyzed by PCR. PCR products were obtained in a final volume of 20 μl containing 40 ng genomic DNA, 300 μM dNTPs, 20 μM for each primer, and 2 units/20 μl of LongAmp Taq DNA Polymerase (New England BioLabs). Primer pairs were designed to be 20–26 bp long for PCR amplification using Primer3 software (Untergasser et al. 2012). The oligonucleotides used are listed in supplementary table

S12, Supplementary Material online. Half of the PCR products were separated on a 1% agarose gel and stained with ethidium bromide for checking the PCR amplification. Fragment sizes were estimated with the 1 kb DNA ladder (Biotools).

Data Access

The SUPERW and SUPERRA tools are available and documented at Sourceforge (https://sourceforge.net/projects/superw/, last accessed July 22, 2015). Illumina paired-end sequences have been deposited in the European Nucleotide Archive SRA and are available at the URL http://www.ebi.ac.uk/ena/data/view/PRJEB7636 (last accessed July 22, 2015).

Supplementary Material

Supplementary material S1, tables S1–S12, and figures S1–S19 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Jordi Morata for his help dating retrotransposon insertions and to Toni Gossmann for comments on a preliminar version of the manuscript. This work was supported by Ministerio de Ciencia e Innovación (grant BFU2009-11932 to J.M.C. and grant AGL2012-40130-C02-01 to J.G.-M.); Ministerio de Economía y Competitividad (grant AGL2013-43244-R to J.M.C., grant AGL2013-41834-R to S.E.R.-O., and grant AGL2009-12698-C02-01 to J.G.-M.); and Fundación Genoma España to J.G.-M.

References

3KRGP, 2014 The 3,000 rice genomes project. *Gigascience* 3:7.

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183:249–258.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet.* 12:363–376.

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

Argyris JM, Ruiz-Herrera A, Madriz Masis P, Sanseverino W, Morata J, Pujol M, Ramos-Onsins SE, Garcia-Mas J. 2015. Use of targeted SNP selection for an improved anchoring of the melon (*Cucumis melo* L.) scaffold genome assembly. *BMC Genomics* 16:4.

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 17:343–360.

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* 24:1242–1255.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* 5:e1000336.

Casacuberta JM, Santiago N. 2003. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene* 311:1–11.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang QY, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods.* 6:677–U676.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet.* 44:803–807.

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345:1249721.

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J.* 63:584–598.

Eduardo I, Arus P, Monforte AJ. 2005. Development of a genomic library of near isogenic lines (NILs) in melon (*Cucumis melo* L.) from the exotic accession PI161375. *Theor Appl Genet.* 112:139–148.

Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, et al. 2013. Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* 25:2783–2797.

El Baidouri M, Panaud O. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol Evol.* 5:954–965.

Esteras C, Formisano G, Roig C, Diaz A, Blanca J, Garcia-Mas J, Gomez-Guillamon ML, Lopez-Sese AI, Lazaro A, Monforte AJ, et al. 2013. SNP genotyping in melons: genetic variation, population structure, and linkage disequilibrium. *Theor Appl Genet.* 126:1285–1303.

Falchi R, Vendramin E, Zanon L, Scalabrin S, Cipriani G, Verde I, Vizzotto G, Morgante M. 2013. Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. *Plant J.* 76:175–187.

Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* 27:343–349.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Felsenstein J. 1978. The number of evolutionary trees. *Syst Zool.* 27:27–33.

Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD. 2012. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol Biol Evol.* 29:675–687.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Gao Z-Y, Zhao S-C, He W-M, Guo L-B, Peng Y-L, Wang J-J, Guo X-S, Zhang X-M, Rao Y-C, Zhang C, et al. 2013. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc Natl Acad Sci U S A.* 110:14492–14497.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E, et al. 2012. The genome of melon (*Cucumis melo* L.). *Proc Natl Acad Sci U S A.* 109:11872–11877.

Garcia-Mas J, Oliver M, Gomez-Paniagua H, de Vicente MC. 2000. Comparing AFLP, RAPD and RFLP markers for measuring genetic diversity in melon. *Theor Appl Genet.* 101:860–864.

Garcia-Mas J, van Leeuwen H, Monfort A, Carmen de Vicente M, Puigdomènech P, Arús P. 2001. Cloning and mapping of resistance gene homologues in melon. *Plant Sci.* 161:165–172.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet.* 8:77–84.

Gebhardt C. 2013. Bridging the gap between genome analysis and precision breeding in potato. *Trends Genet.* 29:248–256.

Godfray HC, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C. 2010. Food security: the challenge of feeding 9 billion people. *Science* 327:812–818.

González VM, Aventín N, Centeno E, Puigdomènech P. 2013. High presence/absence gene variability in defense-related gene clusters of *Cucumis melo*. *BMC Genomics* 14:782.

González VM, Aventín N, Centeno E, Puigdomènech P. 2014. Interspecific and intraspecific gene variability in a 1-Mb region containing the highest density of NBS-LRR genes found in the melon genome. *BMC Genomics* 15:1131.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet.* 45:51–58.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38:e199.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.

Henaff E, Vives C, Desvoyes B, Chaurasia A, Payet J, Gutierrez C, Casacuberta JM. 2014. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of *Brassica* species. *Plant J.* 77:852–862.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26:121–135.

Huang XH, Lu TT, Han B. 2013. Resequencing rice genomes: an emerging new era of rice genomics. *Trends Genet.* 29:225–232.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132:583–589.

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, et al. 2012. Comparative population genomics of maize domestication and improvement. *Nat Genet.* 44:808–811.

Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol.* 27:650–660.

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

Jeffrey C. 1980. A review of the Cucurbitaceae. *Bot J Linn Soc.* 81:233–247.

Jiao YP, Zhao HN, Ren LH, Song WB, Zeng B, Guo JJ, Wang BB, Liu ZP, Chen J, Li W, et al. 2012. Genome-wide genetic changes during modern breeding of maize. *Nat Genet.* 44:812–U124.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* 304:982.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet.* 42:1053–1059.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 15:R84.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet.* 46:1220–1226

Lisch D. 2013. How important are transposons for plant evolution? *Nat Rev Genet* 14:49–61.

Lu C, Chen JJ, Zhang Y, Hu Q, Su WQ, Kuang HH. 2012. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol Biol Evol.* 29:1005–1017.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482:173–178.

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A. 2009. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461:1135–1138.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159:1295–1308.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc Natl Acad Sci U S A.* 110:8615–8620.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet.* 14:840–852.

Monforte AJ, Garcia-Mas J, Arus P. 2003. Genetic variability in melon based on microsatellite variation. *Plant Breed.* 122:153–157.

Morgante M, De Paoli E, Radovic S. 2007. Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol.* 10:149–155.

Myles S. 2013. Improving fruit and wine: what does genomics have to offer? *Trends Genet* 29:190–196.

Olsen KM, Wendel JF. 2013. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu Rev Plant Biol.* 64:47–70.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556.

Pecinka A, Abdelsamad A, Vu GT. 2013. Hidden genetic nature of epigenetic natural variation in plants. *Trends Plant Sci.* 18:625–632.

Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for *Arabidopsis* retrotransposons: a role for targeted integration in the distribution of the Metaviridae. *Genome Biol.* 5:R78.

Pitrat M. 2008. Melon (*Cucumis melo* L.). In: Prohens J, Nuez F, editors. Handbook of crop breeding. Vol. I: Vegetables. New York: Springer. p. 283–315.

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, et al. 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat Genet.* 45:1510–1515.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28:i333–i339.

Saxena RK, Edwards D, Varshney RK. 2014. Structural variations in plant genomes. *Brief Funct Genomics.* 13:296–307

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115.

Sebastian P, Schaefer H, Telford IR, Renner SS. 2010. Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proc Natl Acad Sci U S A.* 107:14269–14273.

Sharma A, Presting GG. 2014. Evolution of centromeric retrotransposons in grasses. *Genome Biol Evol.* 6:1335–1352.

Silberstein L, Kovalski I, Huang R, Anagnostou K, Jahn M, Perl-Treves R. 1999. Molecular variation in melon (*Cucumis melo* L.) as revealed by RFLP and RAPD markers. *Sci Hort.* 79:101–111.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 5:e1000734.

Stepansky A, Kovalski I, Perl-Treves R. 1999. Intraspecific classification of melons (*Cucumis melo* L.) in view of their phenotypic and molecular variation. *Plant Syst Evol.* 217:313–332.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat Genet.* 43:1160–U1164.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tian Z, Zhao M, She M, Du J, Cannon SB, Liu X, Xu X, Qi X, Li M-WW, Lam H-MM, et al. 2012. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. *The Plant Cell* 24:4422–4436.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* 40:e115.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu SQ, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet.* 45:487–U447.

Xu X, Liu X, Ge S, Jensen JD, Hu FY, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol.* 30:105–U157.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25:2865–2871.

Zeng K, Fu Y-XX, Shi S, Wu C-II. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.

Methods

Highly efficient gene tagging in the bryophyte *Physcomitrella patens* using the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon

Cristina Vives[1]*, Florence Charlot[2]*, Corinne Mhiri[2]*, Beatriz Contreras[1], Julien Daniel[2], Aline Epert[2], Daniel F. Voytas[3], Marie-Angèle Grandbastien[2], Fabien Nogué[2] and Josep M. Casacuberta[1]

[1]Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Cerdanyola del Vallès, 08193 Barcelona, Spain; [2]INRA AgroParisTech, IJPB, UMR 1318, INRA centre de Versailles, route de Saint Cyr, 78026 Versailles Cedex, France; [3]Department of Genetics, Cell Biology & Development and Center for Genome Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Authors for correspondence:
Josep M. Casacuberta
Tel: +34 935636600
Email: josep.casacuberta@cragenomica.es

Fabien Nogué
Tel: +33 130833009
Email: fabien.nogue@versailles.inra.fr

Received: *2 June 2016*
Accepted: *13 July 2016*

New Phytologist (2016)
doi: 10.1111/nph.14152

Key words: gene tagging, mutant population, *Physcomitrella patens*, Retrotransposon, Tnt1.

Summary

- Because of its highly efficient homologous recombination, the moss *Physcomitrella patens* is a model organism particularly suited for reverse genetics, but this inherent characteristic limits forward genetic approaches.
- Here, we show that the tobacco (*Nicotiana tabacum*) retrotransposon Tnt1 efficiently transposes in *P. patens*, being the first retrotransposon from a vascular plant reported to transpose in a bryophyte. Tnt1 has a remarkable preference for insertion into genic regions, which makes it particularly suited for gene mutation.
- In order to stabilize Tnt1 insertions and make it easier to select for insertional mutants, we have developed a two-component system where a mini-Tnt1 with a retrotransposition selectable marker can only transpose when Tnt1 proteins are co-expressed from a separate expression unit.
- We present a new tool with which to produce insertional mutants in *P. patens* in a rapid and straightforward manner that complements the existing molecular and genetic toolkit for this model species.

Introduction

Bryophytes were the earliest extant lineage to diverge from the land plant evolutionary lineage *c.* 470 million yr ago. One of their representatives, the moss *Physcomitrella patens*, is a longstanding model for studying plant development, growth and cell differentiation. Interest in this nonvascular plant increased following the discovery that homologous recombination is an efficient process in *P. patens* (Schaefer & Zryd, 1997), which made this species a tool of choice to study gene function in plants. Reverse genetic studies have addressed the relationship between moss organogenesis and many regulators of plant development including growth factors, the cytoskeleton, transduction pathways, transcription factors, epigenetic control and dedifferentiation processes (Bonhomme *et al.*, 2013).

However, the high efficiency of homologous recombination in *P. patens* is Janus faced and can limit forward genetic strategies. Indeed, the extremely low integration efficiency in the absence of any sequence similarity (Schaefer & Zryd, 1997) hampers the development of mutant collections based on *Agrobacterium tumefaciens*-mediated T-DNA insertion.

Recently, and for the first time in *Physcomitrella*, an elegant positional cloning approach using a UV-C based mutant collection allowed the identification of an essential gene for abscisic acid responses in *Physcomitrella* (Stevenson *et al.*, 2016). This approach was made possible thanks to the production of a significant number of molecular markers (Kamisugi *et al.*, 2008) and could potentially be facilitated by tools such as oriented crosses (Perroud *et al.*, 2011). It should be noted that one limitation of this strategy is the need for the isolated mutants to produce functional gametophores. To tackle this limitation, two different mutant collections have been constructed in *P. patens* thanks to

*These authors contributed equally to this work.

the high efficiency of homologous recombination. These two collections rely on transposon-based shuttle mutagenesis systems using genomic DNA (Nishiyama *et al.*, 2000) or cDNA (Egener *et al.*, 2002; Schween *et al.*, 2005). These mutant collections have been screened for different phenotypes and permitted the identification of genes involved in specific physiological process. However, as a consequence of complex and multiple integration events in many tagged lines, probably related to ectopic recombination processes, the link between the observed phenotype and the causal tagged gene was not always straightforward (Hayashida *et al.*, 2005; Schulte *et al.*, 2006). For these reasons, forward genetic approaches are still challenging in *P. patens*. In vascular plants, T-DNA insertion-based strategies as well as alternative approaches based on the insertion of DNA transposons or long terminal repeat (LTR) retrotransposons have been used to produce mutant collections (Sundaresan, 1996). LTR retrotransposons are mobile genetic elements that transpose via an RNA intermediate which is reverse transcribed into DNA by an element-encoded reverse transcriptase. This makes LTR retrotransposons particularly useful as mutagenic agents as their insertions, and the mutations they create, are stable. In addition, some LTR retrotransposons preferentially insert within gene-rich regions (Okamoto & Hirochika, 2000; Le *et al.*, 2007; Urbański *et al.*, 2012). These characteristics make LTR retrotransposons particularly suitable for gene tagging. A number of LTR retrotransposons have been used to generate mutant collections in plants, using both endogenous and heterologous elements. Examples include the rice (*Oryza sativa*) Tos 17 (Hirochika, 2001) and *Lotus japonicus* LORE1 elements (Fukai *et al.*, 2012; Urbański *et al.*, 2012), used in their respective native genomes, as well as the tobacco (*Nicotiana tabacum*) Tnt1 element, used in heterologous hosts such as *Medicago truncatula* (Tadege *et al.*, 2008). Therefore, in an attempt to set up an efficient insertion mutagenesis system in *P. patens*, we explored the possibility of building an insertion mutant collection based on retrotransposon insertions.

We chose the tobacco retrotransposon Tnt1 as this element has been shown to transpose efficiently in different heterologous systems, such as Arabidopsis, *M. truncatula*, lettuce (*Lactuca sativa*) and soybean (*Glycine max*) (Lucas *et al.*, 1995; D'Erfurth *et al.*, 2003; Mazier *et al.*, 2007; Cui *et al.*, 2013).

Here, we show that Tnt1 transposes efficiently in *P. patens* and that it shows a remarkable preference for insertion into genic regions. In order to create a collection of mutants in which the inserted Tnt1 elements could be selected for, we constructed a defective element, which we named mini-Tnt1, containing a selectable marker only active after retrotransposition. In addition, this mini-Tnt1 element is only able to transpose in the presence of Tnt1 proteins expressed from a different plasmid, and therefore the introduced mini-Tnt1 elements will be stabilized and will no longer be able to generate new mutations. We report here the efficient transposition of this mini-Tnt1 element in *P. patens*. In addition to being the first report of the transposition of a retrotransposon from a vascular plant into a bryophyte, our results show that a Tnt1-based insertion mutagenesis system could be an extremely useful tool for forward genetics in *P. patens*.

Materials and Methods

Plant material

Physcomitrella patens (Hedw.) B.S.G. 'Gransden2004' was vegetatively propagated as previously described (Cove *et al.*, 2009). Individual plants were cultured as 'spot inocula' on BCD (1 mM $MgSO_4$, 1.85 mM KH_2PO_4 (pH 6.5, adjusted with KOH), 10 mMKNO$_3$, 45 mM $FeSO_4$, 0.22 mM $CuSO_4$, 0.19 mM $ZnSO_4$,10 mM H_3BO_4, 0.10 mM Na_2MoO_4, 2 mM $MnCl_2$, 0.23 mM $CoCl_2$, 0.17 mM KI) agar medium supplemented with 1 mM $CaCl_2$ and 5 mM ammonium tartrate (BCDAT medium), or as lawns of protonemal filaments by subculture of homogenized tissue on BCDAT agar medium overlaid with cellophane for the isolation of protoplasts.

Bacterial strains and constructs

Plasmid Tnk23 (Lucas *et al.*, 1995) is a derivative of the pBIN19 vector containing the Tnt1-94 retrotransposon element from tobacco (X13777) and a neomycin phosphotransferase (*nptII*) gene that was used as a transformation marker to select for the primary transformants.

The tagged mini-Tnt1 element was constructed as follows. A double-stranded oligonucleotide corresponding to a previously described artificial intron (Hou *et al.*, 2010) was cloned into the *Msc*I site of the pBNRf plasmid containing an *nptII* expression cassette (Schaefer *et al.*, 2010) and a clone, pJCMN21, with the intron in reverse orientation with respect to the *nptII* cassette was selected. The interrupted cassette was amplified by PCR with the oJCBC4 and oJCBC5 primers (Supporting Information Table S1) and was cloned into the pCRII plasmid (Invitrogen, Carlsbad, CA, USA), obtaining pJCBC3. An *Eco*RI fragment of pJCBC3, containing the *nptII* interrupted cassette, was cloned into pENTR3C (Invitrogen). The 5′ fragments of Tnt1 were obtained from pBSX1 (Lucas *et al.*, 1995), by digestion to give the *Sal*I-*Bam*HI (long mini-Tnt1) or *Sal*I-*Bgl*II fragments (short mini-Tnt1), whereas the 3′ fragment was amplified by PCR with the oJCBC6 and oJCBC7 primers (Table S1) on the pBSX1 plasmid. Both fragments were cloned into the pENTR3C plasmid containing the interrupted *nptII* cassette to give the pBC5 (long mini-Tnt1) and pBC7 (short mini-Tnt1) plasmids. The *Xho*I fragments of pBC5 and pBC7, which contain the complete mini-Tnt1 elements, were cloned into the *Xho*I site of the pBHRf vector containing a hygromycin resistance cassette (Schaefer *et al.*, 2010) to obtain the pBC12 (long mini-Tnt1, with an element of 5325 nt) and the pBC11 (short mini-Tnt1, with an element of 3420 nt) plasmids.

For the transient expression of Tnt1 proteins, two types of vector were created. The *Apa*I fragment of a plasmid containing the *nos* terminator was cloned into *Apa*I of pBSX1 (Lucas *et al.*, 1995), eliminating the Tnt1 3′ LTR, to obtain the pJCMN5 plasmid. The vector pBC6 expressing the wild-type Tnt1 proteins under the control of the Tnt1 5′ LTR and *nos* terminator was obtained by cloning the *Sal*I-*Sma*I fragment of pJCMN5 into *Sal*I-*Sma*I sites of the pENTR3C plasmid. Another protein construct was created by mutating the second amino acid, D, in the integrase DDE domain. The mutated integrase, in which D was changed to A, was amplified by PCR using a combination of four primers: oJCMN13, oJCMN14 (which includes the mutation), oJCMN15 (which includes the mutation) and oJCMN16 (Table S1). This fragment was cloned into the PCR8/GW/TOPO cloning vector (Invitrogen) to give the pJCMN4 plasmid. The *Nhe*I-*Nde*I fragment from pJCMN4 was cloned into the *Nhe*I-*Nde*I sites of pJCMN5 to produce the pJCMN7 plasmid. The *Bam*HI-*Nco*I fragment from pJCMN7 was then transferred into the corresponding *Bam*HI-*Nco*I sites of pBC6 to give the pBC10 (mutated proteins) plasmid. These plasmids were digested with *Xho*I and *Nru*I, and the Tnt1 protein-encoding cassette was cloned into the *Xho*I-*Nru*I site of the pBZRf vector, which carries a 35S::zeoR cassette (from the p35S-loxP-Zeo vector; a gift of Prof. M. Hasebe), cloned between two LoxP sites in direct orientation in a pMCS5 backbone (MoBiTec GmbH, Göttingen, Germany). This resulted in plasmids pBC13 (wild-type proteins) and pBC14 (mutated integrase).

Plant transformation and selection

Transformation experiments were performed by protoplast polyethylene glycol fusion as previously described (Trouiller *et al.*, 2006). A total of 4×10^6 protoplasts were transformed with supercoiled DNA of plasmid Tnk23. Aliquots of 10 μg of DNA were used to transform 4×10^5 protoplasts. Protoplasts were plated on cellophane-covered regeneration plates (10^5 protoplasts per plate) containing BCDAT medium with mannitol and incubated in the light (15 W m^{-2}) for 6 d. Antibiotic-resistant plants were selected by transfer of the cellophane overlays for 3 d to BCDAT medium containing G418 (Duchefa Biochem, Haarlem, the Netherlands) (25 μg ml^{-1}) for Tnk23 transformants and hygromycin (Duchefa Biochem) (20 μg ml^{-1}) and zeomycin (20 μg ml^{-1}) (Duchefa Biochem) when appropriate for mini-Tnt1 transformants. For Tnk23 clones, the cellophane overlays were then transferred to BCDAT medium for 10 d and transformed plants were observed and the number of transformed clones was estimated. For mini-Tnt1 transformed plants, the cellophane overlays were transferred to BCDAT medium containing G418 (50 μg ml^{-1}) for 10 d.

PCR analysis of transformants

Moss DNA from Tnk23 and mini-Tnt1 transformants was prepared as previously described (Trouiller *et al.*, 2006). The primers Tnt1-ol13 and Tnt1-Avi1a (Table S1) were used to amplify the LTR of the retrotransposon Tnt1. The primers KanFwd and KanRev were used to amplify a region of the *nptII* gene of the Tnk23 construct. The primers oBC1 and oBC2 were used to amplify almost the entire mini-Tnt1 element. The primers JCMN23 and JCMN24 were used to amplify the flanking region of the intron in reverse orientation with respect to *nptII* of the mini-Tnt1 constructs. The primers oCV1 and oCV2 were used to amplify a region of the hygromycin gene of the mini-Tnt1 plasmids. The primers JCMN15 and JCMN16 were used to amplify a region of the Tnt1 protein plasmids. The primers APTg-f, APT-r, PpAPT#16 and PpAPT#19 were used as a control. The PCR protocol was: 5 min at 95°C, 30 cycles of 40 s at 95°C, 40 s at 60°C, and 1 min at 72°C, and 4 min at 72°C, and storage at 4°C. The amplification product of the intron in reverse orientation with respect to *nptII* was cloned using the TOPO TA Cloning™ kit (Invitrogen) and transformed into *Escherichia coli* strain DH5α by heat shock (Sambrook *et al.*, 1989). Selected clones were grown up and their plasmid DNA was extracted using the Wizard Plus SV Minipreps DNA Purification System™ (Promega, Madison, WI, USA). Clones containing the insert were selected by digestion using *Eco*RI and were sequenced using the universal M13 forward and reverse primers. Sequences of primers used in this study can be found in Table S1. The schematic representation of all plasmids used in this work is shown in Fig. S2.

Real-time RT-PCR

Total RNA was extracted from 7-d-old cultivated protonema using the PureLink™ RNA Mini Kit (Applied Biosystems, Ambion, Foster City, CA, USA). Genomic DNA was eliminated by treatment with the DNA-free™ kit (Applied Biosystems, Ambion). One microgram of total RNA was used to synthesize first-strand cDNA using the SuperScript™ III Reverse Transcriptase kit (Invitrogen).

The quantitative real-time RT-PCR reactions (qRT-PCR) were performed on optical 96-well plates in the Roche LightCycler 480 instrument using SYBR Green I Master (Roche Applied Science, Penzberg, Germany), primers at 10 μM and 1/5 of the cDNA obtained from the reverse transcription of 100 ng of RNA, running each sample in triplicate. The cycling conditions were: 95°C for 5 min (holding stage); then 95°C for 10 s, 56°C for 10 s and 72°C for 10 s (amplification stage); and finally, the qRT-PCR specificity was checked with the melting curve. Reverse transcriptase negative controls and nontemplate controls were included. The adenine phosphoribosyl transferase (*APT*) gene (Schaefer *et al.*, 2010) was used as an internal control to normalize the qRT-PCR output, where Threshold cycle (C_t) values of 40 or above were considered negative values or lack of amplification. Primers were designed using the Primer3 software tool (Untergasser *et al.*, 2012). The primers qAPTf and qAPTr were used to amplify the *APT* transcripts, and the primers qTnt1f and qTnt1r were used to amplify Tnt1 transcripts. Sequences of the primers used in this study can be found in Table S1.

Sequence-specific amplification polymorphism (SSAP) analysis and characterization of SSAP fragments

The SSAP strategy as well as SSAP adaptors and primer sequences used in this study have already been described in Petit *et al.* (2007). Briefly, after *Eco*RI digestion and adaptor primer ligation, genomic DNA was amplified with the E00 adaptor primer and the ^{33}P-labeled Tnt1-ol16 primer located in the LTR of the tobacco Tnt1-94 (X13777) element and orientated towards the 5′ end. Amplified ^{33}P-PCR products were separated on 6% acrylamide gels and exposed after drying to Kodak Biomax™ films (Carestream Health

Inc., Rochester, NY, USA). Only reproducible, intense and clearly identifiable bands were manually scored as present.

SSAP bands of interest were excised from the dried gel with a scalpel and incubated overnight in 30 µl of distilled water at 37°C. Five microliters of each collected eluate was submitted to PCR amplification using the same primers as for SSAP and the following PCR program: 5 min at 94°C, 35 cycles of 30 s at 94°C, 30 s at 56°C and 1 min at 72°C, 10 min at 72°C, and storage at 4°C. After visualization and quantification on a 1% agarose gel, amplification products were cloned using the dual TOPO TA Cloning™ kit (Invitrogen) and used to transform One Shot chemically competent bacteria (Invitrogen) according to the manufacturer's recommendations. Clones containing the inserts were selected by PCR on bacteria using the T7 and SP6 universal primers located on each side of the pCR2.1-TOPO™ cloning vector (Invitrogen). Sequencing of PCR products was performed by Genoscreen (Lille, France) and Beckman Coulter Genomics (Grenoble, France). Sequences of the primers used in this study can be found in Table S1.

Cloning the 3′ junction of newly transposed copies

Primers were designed to bind into the genomic region flanking three Tnt1 insertions (sequences of the primers used in this study can be found in Table S1) and used in combination with the Tnt1-ol16 primer to amplify the 3′ junction. The bands obtained after PCR were cloned and sequenced.

Characterization of 5′ and 3′ Tnt1 junctions

5′ and 3′ flanking sequences from insertions Tnk23-1, Tnk23-8 and Tnk23-13 were PCR amplified using, respectively, primers Tnt1#1 and Tnt2#2 (located in the Tnt1 LTR and oriented outwards from the retrotransposon) in combination with (1) tnk23-1#1 for the 5′ border and tnk23-1#2 for the 3′ border of Tnk23-1; (2) tnk23-8#1 for the 5′ border and tnk23-8#2 for the 3′ border of Tnk23-8; (3) tnk23-13#1 for the 5′ border and tnk23-13#2 for the 3′ border of Tnk23-13. Sequences of the primers used in this study can be found in Table S1.

Results

Tnt1 transposes in *Physcomitrella patens*

In order to investigate the potential transposition of the Tnt1 element in *P. patens*, we transformed *P. patens* protoplasts with the plasmid Tnk23 containing the Tnt1-94 copy (Grandbastien *et al.*, 1989) together with a kanamycin resistance gene (Lucas *et al.*, 1995). After transformation, protoplasts were allowed to recover in nonselective medium for 6 d and selection was applied to select transformed clones for 3 d. Note that this selects for transformed clones and not necessarily for clones that have stably integrated the plasmid, as in *P. patens* plasmids can be maintained in a nonintegrated form for a long time (Ashton *et al.*, 2000; Murén *et al.*, 2009). Resistant clones (*c.* 8000 clones) were further cultivated on nonselective medium for 2 wk. Two hundred clones were then individually transferred in parallel to nonselective or selective medium. All the clones showed a complete loss of the resistance phenotype. Moreover, PCR analysis of 18 independent clones showed that none of them had integrated the *nptII* gene (Fig. 1, middle panel). This was expected because the integration of nonhomologous sequences in *P. patens* is extremely inefficient, the mean relative transformation frequency (RTF) of nonhomologous supercoiled plasmids being close to 1×10^{-5} (Schaefer & Zryd, 1997). Interestingly, the PCR analysis of the selected clones showed that all of them had a Tnt1 sequence(s) integrated (Fig. 1, top panel). The absence of integrated plasmid sequences such as the *nptII* gene in the selected *P. patens* clones (Fig. 1, middle panel) strongly suggests that Tnt1 inserted into the genome after retrotransposition from nonintegrated plasmids.

To confirm Tnt1 retrotransposition and determine the number of Tnt1 sequences integrated in each selected clone, we performed SSAP experiments on the same 18 transformed *P. patens* clones to amplify Tnt1/genome junctions using the Tnt1-ol16

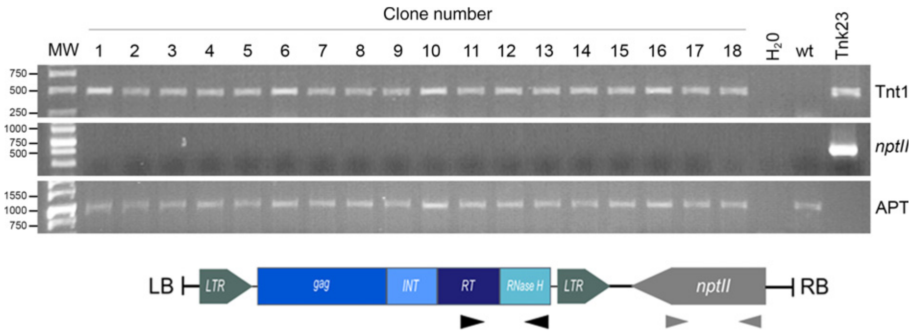


Fig. 1 Analysis of the presence of Tnt1 and neomycin phosphotransferase (*nptII*) sequences in 18 transformed *Physcomitrella patens* clones. PCR amplification of Tnt1 (top panel), *nptII* (middle panel) and the endogenous *P. patens* adenine phosphoribosyl transferase (*APT*) (bottom panel) gene sequences in the *P. patens* clones transformed with the Tnk23 plasmid. Information on the DNA and control used, including the number of *P. patens* clones, is shown at the top. A diagram displaying the structure of the Tnk23 T-DNA sequences together with the approximate position of the primers used to amplify Tnt1 and *nptII* sequences is shown below. The two long terminal repeats (LTRs) of the Tnt1 element are shown in grey and the different proteins it encodes are shown in the blue shaded boxes: gag; INT, integrase; RT, reverse transcriptase; RNase H.

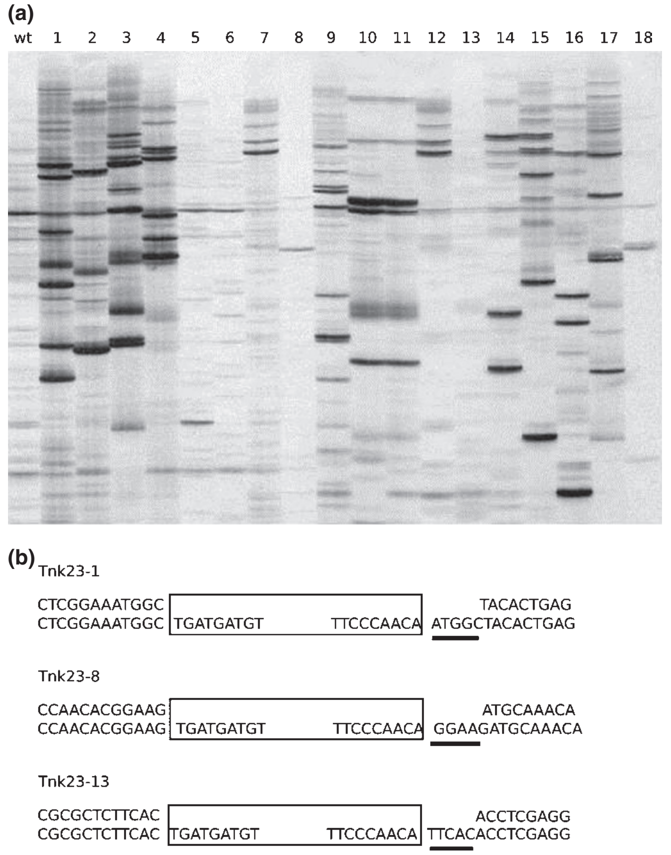

(a)
wt 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
(b)
Tnk23-1
CTCGGAAATGGC TACACTGAG
CTCGGAAATGGC TGATGATGT TTCCCAACA ATGGCTACACTGAG
Tnk23-8
CCAACACGGAAG ATGCAAACA
CCAACACGGAAG TGATGATGT TTCCCAACA GGAAGATGCAAACA
Tnk23-13
CGCGCTCTTCAC ACCTCGAGG
CGCGCTCTTCAC TGATGATGT TTCCCAACA TTCACACCTCGAGG

Fig. 2 Analysis of Tnt1 insertions. (a) Sequence-specific amplification polymorphism (SSAP) analysis of 18 different *Physcomitrella patens* clones transformed with the Tnk23 plasmid (the corresponding number is given at the top) as well as an untransformed *P. patens* clone (wt). (b) Sequence of three Tnt1 insertion junctions. The 12-nt 5′ and 14-nt 3′ flanking sequences are shown together with the first and last 9 nt of Tnt1 (boxed). A comparison with the corresponding pre-insertion locus is shown above. The target site duplications are underlined. The Tnt1 insertion name is indicated above each diagram. The complete 5′ flanking sequence of these three insertions are given in Supporting Information Fig. S1.

primer designed based on the LTR of Tnt1. The SSAP analysis shown in Fig. 2(a) indicates that the number of Tnt1 elements inserted in each clone varied from 1 to *c.* 10.

In order to determine whether Tnt1 inserted into the genome through retrotransposition, we cloned and sequenced 22 SSAP bands, and the analysis of these sequences showed that in all cases the inserted Tnt1 sequence started with the first nucleotide of the 5′ LTR of Tnt1 and no plasmid sequence was inserted (Fig. S1). Moreover, the additional analysis of the corresponding 3′ junctions for three Tnt1 insertions showed that the inserted sequence ended precisely at the last nucleotide of the 3′ LTR (Fig. 2b). An analysis of the sequences flanking the insertions revealed in all cases a duplication of 5 nt at the insertion site (Fig. 2b), which is consistent with the target site duplication (TSD) that Tnt1 generates upon retrotransposition (Grandbastien *et al.*, 1989; Lucas *et al.*, 1995). All these data demonstrate that Tnt1 inserted into the genome of *P. patens* by retrotransposition. The absence, in all analyzed cases, of plasmid sequence flanking Tnt1 elements, together with the absence of *nptII* sequences in all tested clones, confirms that Tnt1 elements transposed from the nonintegrated Tnk23 plasmid transiently maintained in the transformed protoplasts. To our knowledge, this is the first report of the transposition of a retrotransposon from a vascular plant into a bryophyte, and opens up the possibility of using the tobacco Tnt1 retrotransposon as an insertional mutagen in *P. patens*.

Tnt1 targets genic regions for insertion

Tnt1 inserts preferentially in genic regions in its natural host (Le *et al.*, 2007). Furthermore, Tnt1 seems to maintain this preference in a heterologous host. Indeed, when introduced to Arabidopsis, *M. truncatula* and soybean it inserts within genes in 73%, 30% and 66% of cases, respectively, while these sequences only account for 45%, 15% and 10% of the genome, respectively (Courtial *et al.*, 2001; D'Erfurth *et al.*, 2003; Cui *et al.*, 2013). This preferential insertion within genes would allow a reduction of the number of independent insertions needed to saturate the genome with genic insertions. We thus analyzed the sequences flanking 22 Tnt1 insertions in different *P. patens* clones. The 22 independent loci analyzed are located on 18 different chromosomes (Table 1), showing that Tnt1 transposes throughout the genome of *P. patens*. In 54% of the cases, Tnt1 inserted within a known protein-coding gene (Table 1). As annotated genes account only for 17% of the genome of *P. patens* (Rensing *et al.*, 2008), these data show that Tnt1 has a strong bias for insertion into genes in *P. patens*. Five other insertions (22%) lie at < 1000 nt from a known coding region, increasing the percentage of insertions potentially compromising gene activity to 77%. This target site preference for gene regions is not very different from that reported for Tnt1 in other plant genomes, showing that the genome of *P. patens*, which is a haploid genome phylogenetically very distant from the genomes of the dicotyledonous plants tested so far, does not behave differently with respect to the insertion pattern of a retrotransposon such as Tnt1.

The inserted Tnt1 elements maintain their capacity to transcribe and potentially transpose

In tobacco, its natural host, Tnt1 is an active element whose expression can only be detected in protoplasts and in other defense-related situations (Grandbastien *et al.*, 2005). In order to analyze whether Tnt1 copies stably integrated in the genome of *P. patens* are expressed, we analyzed the transcription of Tnt1 in protonema tissue as well as in protoplasts. The analysis of the expression of two independent Tnt1 clones is presented in Fig. 3(a). This analysis shows that the newly integrated Tnt1 elements were expressed in protonema tissue and that, in sharp contrast to what happens in its natural host and in heterologous flowering plant species where it has been introduced, its expression was reduced in protoplasts. Interestingly, when protoplasts were cultivated and allowed to reform filaments, Tnt1 expression was recovered, confirming the specific and transient decrease of expression associated with the production of protoplasts. It has been recently shown that different retrotransposon families are expressed in protonema tissue in *P. patens* (S. Rensing, pers.

Table 1 Analysis of sequences flanking Tnt1 and mini-Tnt1 insertions. The Tnk23 construct corresponds to the full Tnt1 tobacco retrotransposon, while the mini-Tnt1 constructs are pBC12 (long) and pBC11 (short)

Tnt1 or mini-Tnt1 insertion	Chromosome number	Tnt1 position	Closest gene (Phytozome and Cosmoss numbers)	Coordinates of closest gene (ATG to stop)	Distance to ATG or stop	Joint Genome Institute (JGI) annotation of closest gene
tnk23-1	Chr_18	14998569	Pp3c18_21350 Pp1s33_94V6.1	14998680–15001467	111 bp from ATG	Alpha/beta-Hydrolase related
tnk23-2	Chr_25	9806656	Pp3c25_13850 Pp1s57_226V6.1	9803950–9807740	Between ATG and STOP Exon 7	Transcription factor Transparent Testa 8 (TT8)
tnk23-3	Chr_12	2583945	Pp3c12_3220 Pp1s404_28V6.1	2584445–2586373	500 bp from ATG	Glycosil transferase
tnk23-4	Chr_7	10651220	Pp3c7_15580 Pp1s153_98V6.1	10650230–10654421	Between ATG and STOP Exon 3	Protein Regulator of Cytokinesis 1 (PCR1)
tnk23-5	Chr_8	2023598	Pp3c8_3890 Pp1s35_271V6.1	2030464–2024739	1141 bp from stop	Ikappab Kinase complex-associated protein
tnk23-6	Chr_5	11458333	Pp3c5_16440 Pp1s116_29V6.1	11460023–11458275	Between ATG and STOP Exon 2	No functional annotations
tnk23-7	Chr_3	24860166	Pp3c3_37190 Pp1s96_125V6.1	24867875–24864633	4467 bp from stop	Malaria antigen-related
tnk23-8	Chr_21	4149245	Pp3c21_6880 Pp1s27_371V6.1	4148574–4147201	671 bp from ATG	Domain of unknown function (DUF309)
tnk23-9	Chr_20	1295616	Pp3c20_2380 Pp1s152_115V6.1	1299377–1301373	3761 bp from ATG	Cytokinin dehydrogenase
tnk23-10	Chr_15	2895324	Pp3c15_4590 Pp1s83_113V6.1	2895214–2895393	Between ATG and STOP Exon 1	No functional annotations
tnk23-11	Chr_16	10436260	Pp3c16_17030 Pp1s197_15V6.1	10436682–10435815	Between ATG and STOP Exon 1	Tetraspanin-18-related
tnk23-12	Chr_17	13700803	Pp3c17_20510 Pp1s65_281V6.1	13699939–13697860	864 bp from ATG	Acireductone Dioxygenase
tnk23-13	Chr_27	2745053	Pp3c27_4760 Pp1s54_8V6.1	2744524–2749844	Between ATG and STOP Exon 1	Mechanosensitive ion channel
tnk23-14	Chr_15	1956943	Pp3c15_3360 Pp1s124_156V6.1	1948450–1937321	8493 bp from ATG	Raptor/Kontroller Of Growth 1 (KOG1) homolog
tnk23-15	Chr_16	9192950	Pp3c16_14670 Pp1s15_76V6.1	9192738–9193277	Between ATG and STOP Exon 1	No functional annotations
tnk23-16	Chr_2	6872328	Pp3c2_9950 Pp1s84_139V6.1	6869908–6866047	2420 pb from ATG	No functional annotations
tnk23-17	Chr_23	12616131	Pp3c23_19350 Pp1s49_168V6.1	12615128–12613085	1003 pb from ATG	Tryptophan Permease
tnk23-18	Chr_8	822597	Pp3c8_1720 Pp1s35_57V6.1	822965–821185	Between ATG and STOP Intron1	Glycosyl transferase family 2
tnk23-19	Chr_15	8948196	Pp3c15_13470 Pp1s104_216V6.1	8947882–8949194	Between ATG and STOP Exon1	Phospholipase A2 family protein
tnk23-20	Chr_24	10916803	Pp3c24_16740 Pp1s323_28V6.1	10917349–10916735	Between ATG and STOP Exon 1	Cyclophilin
tnk23-21	Chr_10	6685306	Pp3c10_9990 Pp1s58_281V6.1	6684783–6686696	Between ATG and STOP Intron1	No functional annotations
tnk23-22	Chr_1	753331	Pp3c1_35620 Pp1s28_202V6.1	747391–762757	Between ATG and STOP Intron4	Kinesin-like protein
pBC12-6	Chr_23	7136546	Pp3c23_10200 Pp1s10_17V6.1	7137132–7134014	Between ATG and STOP Exon 1	Auxin efflux carrier component 3-related
pBC11-1	Chr_7	10718730	Pp3c7_15700 Pp1s153_79V6.1	10717393–10719080	Between ATG and STOP Exon 4	F-box-like
pBC12-30	Chr_18	7691183	Pp3c18_10870 Pp1s19_291V6.1	7688119–7695553	Between ATG and STOP Exon 11	Protein tyrosine kinase
pBC12-14	Chr_9	3889027	Pp3c9_6830 Pp1s220_62V6.1	3889604–3892995	577 pb from ATG	No functional annotations

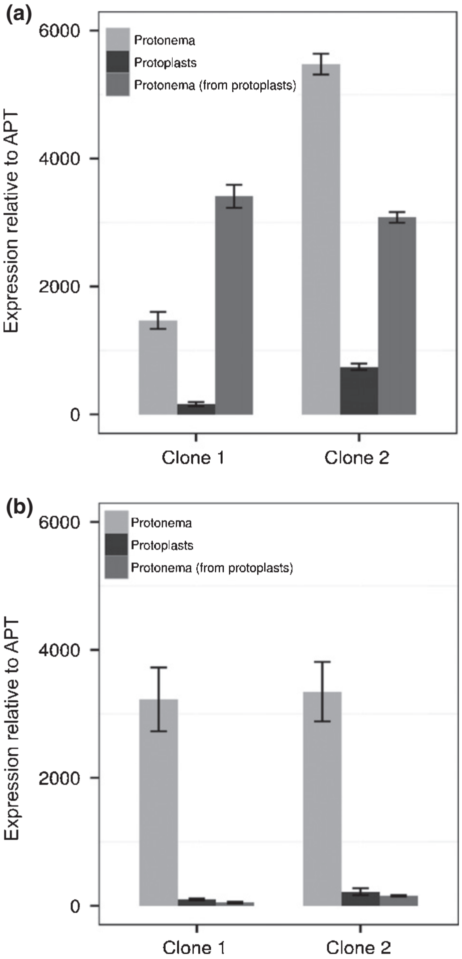


Fig. 3 Expression analysis of (a) Tnt1 and (b) RLG1 retrotransposon families in different cell types of two *Physcomitrella patens* clones transformed with the Tnk23 plasmid. Error bars represent ± SE of three technical replicates. APT, adenine phosphoribosyl transferase.

comm.). Therefore, we decided to analyze the expression of the most highly expressed retrotransposon family, the gypsy RLG1 family, in the same analyzed samples. Fig. 3(b) clearly shows that, whereas RLG1 was highly expressed in protonema, its expression was greatly reduced in protoplasts. This result suggests that the regulation of the tobacco Tnt1 in *P. patens* shows similarities with that of the endogenous elements. More analyses will be required to determine to what extent these differences are indicative of a more profound difference with respect to most plants in the way in which *P. patens* deals with stress and transposon regulation. However, irrespective of the underlying reasons, the pattern of Tnt1 expression in *P. patens* has an impact on the potential use of Tnt1 to generate mutants in this species. Indeed, the expression and potential transposition of Tnt1 in protonema filaments may induce subsequent mutations once a mutant phenotype has been obtained, leading to unstable phenotypes which could make it challenging to establish a link between phenotype and genotype.

For this reason, we decided to design a Tnt1 two-component system separating the mobile unit from the sequences needed to express the proteins required for its retrotransposition.

A Tnt1-based two-component system for efficient mutagenesis in *Physcomitrella patens*

In addition to autonomous transposable elements, genomes contain defective elements that, although not able to autonomously transpose, can be mobilized in *trans* by related elements. Experiments in tobacco protoplasts have shown that mini-Tnt1 elements devoid of the sequences coding for the different proteins required for transposition (i.e. gag and pol, containing the integrase, protease, reverse transcriptase and RNAse H) can be mobilized in *trans* (Hou *et al.*, 2010). Based on these data, we constructed different mini-Tnt1 elements where a variable fraction of the coding sequence was replaced by a retrotransposition marker similar to previously reported ones (Hou *et al.*, 2010; Fig. 4). The presence of an intron in reverse orientation with respect to the resistance gene, but direct orientation with respect to the Tnt1 promoter, ensures that resistance is achieved only after Tnt1 transcription and retrotransposition (Fig. 4). As a protein donor, we used a Tnt1 devoid of the 3′ LTR, which makes it unable to transpose, that we replaced with a conventional transcriptional terminator. As a control, we constructed a vector expressing a defective version of Tnt1 proteins which contains a mutation in the integrase core domain that blocks integration (Ke & Voytas, 1999; Fig. 4). We transformed *P. patens* with different combinations of plasmids to check for Tnt1 retrotransposition. As shown in Table 2, transforming *P. patens* with a plasmid containing a mini-Tnt1 element together with a plasmid expressing the wild-type version of the Tnt1 proteins resulted in *P. patens* clones resistant to kanamycin, whereas no resistant clone was obtained in transformations where only the mini-Tnt1 was used. These results suggest a retrotransposition of the selectable mini-Tnt1 elements. Interestingly, when the mini-Tnt1 plasmid was transformed together with the mutated version of the Tnt1 proteins, no resistant clones were obtained (Table 2). This result further suggests that the mini-Tnt1 elements integrated into the *P. patens* genome through retrotransposition and not by direct integration. It is interesting to note that a much higher number of resistant clones was obtained with the longer version of the mini-Tnt1 compared with the shorter mini-Tnt1 version (34 verus two; see Table 2), which suggests that some internal Tnt1 sequences may be required for efficient transposition. Alternatively, length constraints for integration could also explain the higher efficiency of the long mini-Tnt1 as compared with the shorter version. Further experiments are needed to investigate these possibilities.

An analysis of the resistant clones using PCR showed that the expected size for the mini-Tnt1 sequences could be amplified in all cases (Fig. 5a). In addition, PCR analysis, as well as the sequencing of the amplified band, showed that, in all cases, the *nptII* gene contained in the integrated mini-Tnt1 elements had lost the intron, confirming that the integrated mini-Tnt1 elements had arisen from the reverse transcription of a spliced mini-Tnt1 transcript (Fig. 5b). An analysis using SSAP of 14 clones showed that the copy number ranged between one and two copies per clone (Fig. 6). The cloning and sequencing of four SSAP bands showed that in all cases the Tnt1 sequences started

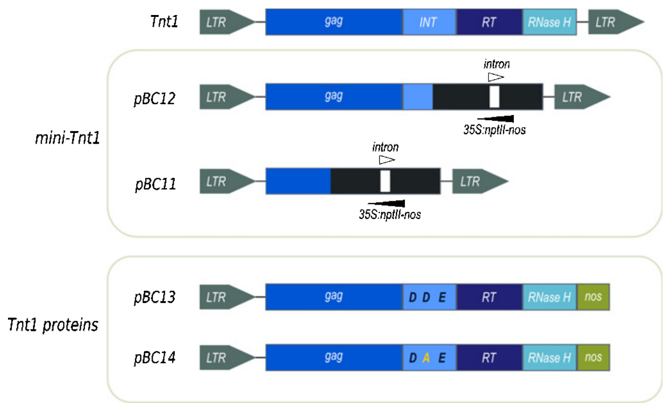


Fig. 4 Tnt1-based two-component retrotransposon system. A variable fraction of coding Tnt1 sequence was replaced by a retrotransposition marker to obtain the long mini-Tnt1 element (pBC12) and the short mini-Tnt1 element (pBC11). The different proteins required for transposition are given in *trans*, where the 3′ long terminal repeat (LTR) is replaced by the *nos* terminator. The pBC13 construct contains the wild-type version of Tnt1 proteins while the pBC14 has a mutation in the integrase core domain. The LTRs of the Tnt1 element are shown in grey and the different proteins it encodes are shown in the blue shaded boxes: gag; INT, integrase; RT, reverse transcriptase; RNase H.

Table 2 Co-transformation results for mini-Tnt1 and protein constructs in *Physcomitrella patens*

Mini-Tnt1 constructs	Protein constructs	No. of transformed protoplasts	No. of KanR clones
pBC12 (long)	pBC13 (wt)	2108	34
pBC12 (long)	pBC14 (mutated)	1576	0
pBC12 (long)	None	3292	0
pBC11 (short)	pBC13 (wt)	2304	2

with the first nucleotide of the 5′ LTR and no plasmid sequence was inserted, which further confirmed that the mini-Tnt1 elements inserted through retrotransposition (Fig. 1c). The analysis of the sequences flanking the inserted mini-Tnt1 elements also confirmed the striking insertion preference of Tnt1 in genic regions in *P. patens*. Indeed, the four analyzed mini-Tnt1 elements inserted in annotated genes (Table 1).

All these results show that the mini-Tnt1 elements efficiently transpose in *P. patens*, targeting genic regions for insertion. Therefore, the Tnt1 two-component system we describe here should allow efficient transposition of selectable mini-Tnt1 elements into the genome of *P. patens*.

Discussion

The high rate of homologous recombination in *P. patens*, which greatly facilitates reverse genetics approaches, makes it a unique model organism for plant research. However, as integration of foreign DNA with no sequence similarity is inefficient, forward genetic analyses are very difficult in this species. We present here a highly efficient system to create insertional mutants in *P. patens* based on the transposition of the tobacco Tnt1 retrotransposon.

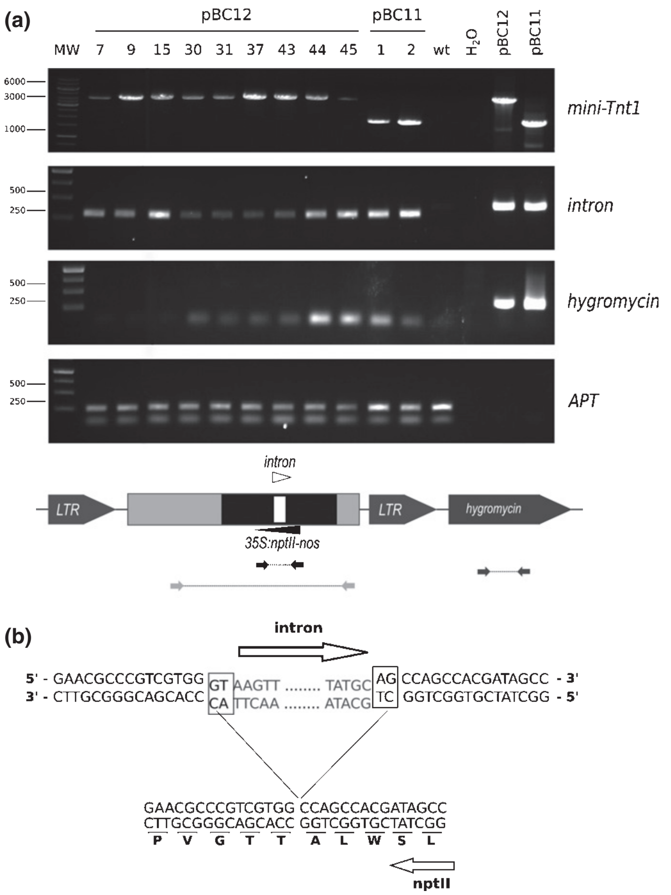

(a)
pBC12
pBC11
MW
7
9
15
30
31
37
43
44
45
1
2
wt
H2O
pBC12
pBC11
6000
3000
1000
mini-Tnt1
500
250
intron
500
250
hygromycin
500
250
APT
intron
LTR
35S:nptII-nos
LTR
hygromycin
(b)
intron
5' - GAACGCCCGTCGTGG GT AAGTT .... TATGC AG CCAGCCACGATAGCC - 3'
3' - CTTGCGGGCAGCACC CA TTCAA .... ATACG TC GGTCGGTGCTATCGG - 5'
GAACGCCCGTCGTGG CCAGCCACGATAGCC
CTTGCGGGCAGCACC GGTCGGTGCTATCGG
P V G T T A L W S L
nptII

Fig. 5 Analysis of the presence of mini-Tnt1 and other regions of the construct. (a) Ten transformed *Physcomitrella patens* clones were analyzed, two of them being transformed with pBC11 (short mini-Tnt1 construct) and eight with pBC12 (long mini-Tnt1 construct). (b) Splicing of the intron sequence. The band amplified with neomycin phosphotransferase (*nptII*) primers flanking the intron was cloned and sequenced. The sequence of the amplified product is shown below the sequence of the plasmid *nptII* gene, showing that the intron sequence was correctly spliced. *APT*, adenine phosphoribosyl transferase.

It was previously proposed that Tnt1 could transpose from nonintegrated plasmids, as transformation of *M. truncatula* with a Tnt1-containing plasmid using *Agrobacterium tumefaciens* resulted in Tnt1-containing calli which in 25% of cases did not contain the corresponding T-DNA sequences (D'Erfurth *et al.*, 2003). This opened up the possibility of using Tnt1 in a system such as *P. patens* where integration of plasmid sequence showing no homology is highly inefficient. The results presented here show that, indeed, both Tnt1 and a mini-Tnt1 element containing a selectable marker of retrotransposition efficiently transpose from nonintegrated plasmids into the genome of *P. patens*.

Tnt1 is only expressed in protoplasts and under defense-related stress situations in tobacco and in all plants where it has been introduced (D'Erfurth *et al.*, 2003; Grandbastien *et al.*, 2005; Mazier *et al.*, 2007; Tadege *et al.*, 2008; Cui *et al.*, 2013). In contrast, our results show that Tnt1 is highly expressed in protonema tissue, whereas its expression decreases in protoplasts, and expression is recovered when protoplasts are cultured and allowed to form new protonema tissues. Tnt1 expression in tobacco is

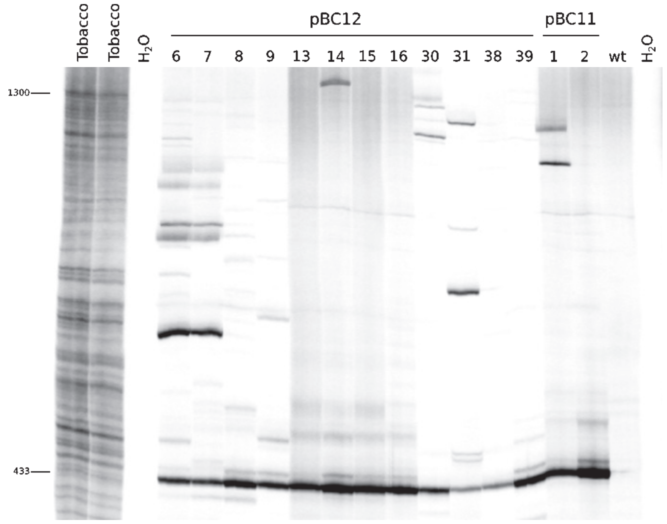


Fig. 6 Sequence-specific amplification polymorphism (SSAP) analysis of mini-Tnt1 insertions in *Physcomitrella patens* clones either transformed with pBC11 (short mini-Tnt1 construct: 12 clones) or with pBC12 (long mini-Tnt1 construct: two clones) together with the pBC13 plasmid harboring the Tnt1 wild-type (wt) proteins necessary for achievement of the retrotransposition process.

induced in stress situations as a consequence of the presence in its promoter of stress-responsive sequences (Beguiristain *et al.*, 2001; Grandbastien *et al.*, 2005). The lack of transcriptional activation of Tnt1 in *P. patens* protoplasts could be attributable to the absence of transcription factors equivalent to those responsible for Tnt1 activation in tobacco, able to bind to the Tnt1 promoter and activate it in this species. However, its high transcription in filaments and the specific and transient reduction of Tnt1 expression in protoplasts suggest a more complex scenario. The fact that the endogenous *P. patens* retrotransposons of the RLG1 group are also expressed in protonema and show a similar decrease of expression in protoplasts suggests important differences with respect to most plants in dealing with stress and transposon regulation in this moss. However, more experiments will be required to address this interesting issue.

In the framework of the present study, this particular pattern of expression has prompted us to develop a two-component system in which the mobile Tnt1 unit would be stabilized (i.e. not able to generate new insertions) after integration in the genome. The mini-Tnt1 two-component system presented here separates the mobile unit, a mini-Tnt1 element which does not encode the proteins needed for transposition, and an element encoding the Tnt1 gag and pol polyproteins which is not mobile because it lacks one of the LTRs needed for mobilization. The mini-Tnt1 unit also contains a selectable marker which is only active after retrotransposition and allows selection of the *P. patens* cells where the mini-Tnt1 has retrotransposed and integrated. Our results show that this two-component system enables the efficient retrotransposition of the mini-Tnt1 elements and that these events can be easily selected in the appropriate medium. Although the inserted mini-Tnt1 elements can still be transcribed in protonema, the insertions are stabilized as the mini-Tnt1 elements are only mobile in the presence of the Tnt1 proteins which are not integrated into the genome, which makes the potential phenotypes created by the insertions stable.

In addition to the interesting characteristics summarized above, the Tnt1-derived insertional mutagenesis system described has the important advantage of targeting preferentially genic regions. Indeed, our results show that 65% of the Tnt1 insertions analyzed lie < 1 kb from an annotated coding region, and the analysis of a small number of mini-Tnt1 insertions, where all four also lie < 1 kb from an annotated coding region, with three of them sitting within a gene (Table 1), confirms this marked preferential insertion into genes. This insertion preference, which Tnt1 shares with other copia-like retrotransposons from plants such as Tto1 or Tos17 (Okamoto & Hirochika, 2000; Miyao *et al.*, 2003), may suggest a tendency for insertion into open chromatin. However, this tendency does not limit Tnt1 insertion to genes expressed in protoplasts and protonema, as the analysis of the expression patterns of the genes where Tnt1 has landed shows that most of them are not preferentially expressed in those tissues (Fig. S3). Tnt1 preferential insertion greatly reduces the total number of independent insertions needed to mutate all 36 000 *P. patens* genes, which in the case of random insertion can be estimated at close to 600 000, applying $P = 1 - (1 - g/G)^n$, where g is the mean size of a gene and G the genome size (Hirochika *et al.*, 2004). In a standard protoplast transformation, up to 50 000 independent resistant clones can be obtained (Schaefer *et al.*, 1991), allowing a complete mutant population to be obtained in just a few days. This ease of mutant selection will also permit the production of mutant population in specific genetic backgrounds, such as *P. patens* lines containing particular markers (Nakaoka *et al.*, 2012) or even suppressor screens in already characterized mutants (St Johnston, 2002). However, it should be noted that the haploid nature of *P. patens* could be a limitation to forward genetic screens when the result of the mutation is detrimental to development. For this reason, and taking into consideration the size of the population needed for saturation, forward genetic screens that can be considered are essentially resistance of the mutants to biotic or abiotic stresses or to drugs (including hormones) that will inhibit the development of the wild-type *P. patens.*

Finally, and in contrast to the existing gene-tagging techniques previously used in *P. patens* (Hayashida *et al.*, 2005; Schulte *et al.*, 2006), the Tnt1 strategy, where the insertions are catalyzed by the retrotransposition machinery, does not involve the homologous recombination machinery of the host. For this reason, the risk of complex and multiple integration events, potentially related to ectopic recombination processes, is abolished using this strategy and the structure of the integration products is highly predictable.

In summary, we present here a new tool to produce insertional mutants in *P. patens* in a rapid and straightforward manner that complements the existing molecular and genetic toolkit for this model species. Together with the fact that *P. patens* is a haploid plant, this will make forward genetics a very efficient tool in this model species and should facilitate the deciphering of the major developmental innovations that were associated with the colonization of land by plants. In particular, this new strategy should contribute to the assignation of function for the numerous genes that are still of unknown function. Finally, the setting up of this forward genetic tool in other model bryophytes, where reverse genetic analysis is not necessarily easy, such as *Ceratodon purpureus* (Trouiller *et al.*, 2007), would be potentially of great interest.

Acknowledgements

Work done at CRAG was supported by the Ministerio de Ciencia y Innovación (grants: BFU2009-11932 and AGL2013-43244-R). We thank Mitsuyasu Hasebe (National Institute for Basic Biology, Okazaki, Japan) for providing the p35S-loxP-Zeo vector.

Author contributions

J.M.C. and F.N. designed the research; C.V., F.C. and C.M. performed the research with the help of B.C., J.D. and A.E. and under the supervision of J.M.C., F.N. and M-A.G.; D.F.V. contributed material and hosted J.C., enabling some constructs to be obtained in his laboratory, and J.C. and F.N. wrote the manuscript with contributions from all the authors.

References

Ashton NW, Champagne CEM, Weiler T, Verkoczy LK. 2000. The bryophyte *Physcomitrella patens* replicates extrachromosomal transgenic elements. *New Phytologist* **146**: 391–402.

Beguiristain T, Grandbastien MA, Puigdomènech P, Casacuberta JM. 2001. Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. *Plant Physiology* **127**: 212–221.

Bonhomme S, Nogué F, Rameau C, Schaefer DG. 2013. Usefulness of *Physcomitrella patens* for studying plant organogenesis. *Methods in Molecular Biology* **959**: 21–43.

Courtial B, Feuerbach F, Eberhard S, Rohmer L, Chiapello H, Camilleri C, Lucas H. 2001. Tnt1 transposition events are induced by *in vitro* transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. *Molecular and General Genetics* **265**: 32–42.

Cove DJ, Perroud P-F, Charron AJ, McDaniel SF, Khandelwal A, Quatrano RS. 2009. Culturing the moss *Physcomitrella patens*. *Cold Spring Harbor Protocols* **2009**: pdb.prot5136.

Cui Y, Barampuram S, Stacey MG, Hancock CN, Findley S, Mathieu M, Zhang Z, Parrott WA, Stacey G. 2013. Tnt1 retrotransposon mutagenesis: a tool for soybean functional genomics. *Plant Physiology* **161**: 36–47.

D'Erfurth I, Cosson V, Eschstruth A, Lucas H, Kondorosi A, Ratet P. 2003. Efficient transposition of the Tnt1 tobacco retrotransposon in the model legume *Medicago truncatula*. *Plant Journal* **34**: 95–106.

Egener T, Granado J, Guitton M-C, Hohe A, Holtorf H, Lucht JM, Rensing SA, Schlink K, Schulte J, Schween G *et al.* 2002. High frequency of phenotypic deviations in *Physcomitrella patens* plants transformed with a gene-disruption library. *BMC Plant Biology* **2**: 6.

Fukai E, Soyano T, Umehara Y, Nakayama S, Hirakawa H, Tabata S, Sato S, Hayashi M. 2012. Establishment of a *Lotus japonicus* gene tagging population using the exon-targeting endogenous retrotransposon LORE1. *Plant Journal* **69**: 720–730.

Grandbastien MA, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa APP, Le QH, Melayah D, Petit M, Poncet C *et al.* 2005. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic and Genome Research* **110**: 229–241.

Grandbastien MA, Spielmann A, Caboche M. 1989. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature* **337**: 376–380.

Hayashida A, Takechi K, Sugiyama M, Kubo M, Itoh RD, Takio S, Fujita T, Hiwatashi Y, Hasebe M, Takano H. 2005. Isolation of mutant lines with decreased numbers of chloroplasts per cell from a tagged mutant library of the moss *Physcomitrella patens*. *Plant Biology* **7**: 300–306.

Hirochika H. 2001. Contribution of the Tos17 retrotransposon to rice functional genomics. *Current Opinion in Plant Biology* **4**: 118–122.

Hirochika H, Guiderdoni E, An G, Hsing YI, Eun MY, Han CD, Upadhyaya N, Ramachandran S, Zhang Q, Pereira A *et al.* 2004. Rice mutant resources for gene discovery. *Plant Molecular Biology* **54**: 325–334.

Hou Y, Rajagopal J, Irwin PA, Voytas DF. 2010. Retrotransposon vectors for gene delivery in plants. *Mobile DNA* **1**: 1–9.

Kamisugi Y, von Stackelberg M, Lang D, Care M, Reski R, Rensing SA, Cuming AC. 2008. A sequence-anchored genetic linkage map for the moss, *Physcomitrella patens*. *Plant Journal* **56**: 855–866.

Ke N, Voytas DF. 1999. cDNA of the yeast retrotransposon Ty5 preferentially recombines with substrates in silent chromatin. *Molecular and Cellular Biology* **19**: 484–494.

Le QH, Melayah D, Bonnivard E, Petit M, Grandbastien MA. 2007. Distribution dynamics of the Tnt1 retrotransposon in tobacco. *Molecular Genetics and Genomics* **278**: 639–651.

Lucas H, Feuerbach F, Kunert K, Grandbastien MA, Caboche M. 1995. RNA-mediated transposition of the tobacco retrotransposon Tnt1 in *Arabidopsis thaliana*. *EMBO Journal* **14**: 2364–2373.

Mazier M, Botton E, Flamain F, Bouchet J-P, Courtial B, Chupeau M-C, Chupeau Y, Maisonneuve B, Lucas H. 2007. Successful gene tagging in lettuce using the Tnt1 retrotransposon from tobacco. *Plant Physiology* **144**: 18–31.

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H. 2003. Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15**: 1771–1780.

Murén E, Nilsson A, Ulfstedt M, Johansson M, Ronne H. 2009. Rescue and characterization of episomally replicating DNA from the moss *Physcomitrella*. *Proceedings of the National Academy of Sciences, USA* **106**: 19444–19449.

Nakaoka Y, Miki T, Fujioka R, Uehara R, Tomioka A, Obuse C, Kubo M, Hiwatashi Y, Goshima G. 2012. An inducible RNA interference system in *Physcomitrella patens* reveals a dominant role of augmin in phragmoplast microtubule generation. *Plant Cell* **24**: 1478–1493.

Nishiyama T, Hiwatashi Y, Sakakibara I, Kato M, Hasebe M. 2000. Tagged mutagenesis and gene-trap in the moss, *Physcomitrella patens* by shuttle mutagenesis. *DNA Research* **7**: 9–17.

Okamoto H, Hirochika H. 2000. Efficient insertion mutagenesis of Arabidopsis by tissue culture-induced activation of the tobacco retrotransposon Tto1. *Plant Journal* **23**: 291–304.

Perroud PF, Cove DJ, Quatrano RS, Mcdaniel SF. 2011. An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytologist* **191**: 301–306.

Petit M, Lim KY, Julio E, Poncet C, Dorlhac de Borne F, Kovarik A, Leitch AR, Grandbastien M-A, Mhiri C. 2007. Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). *Molecular Genetics and Genomics* **278**: 1–15.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y *et al.* 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**: 64–69.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: a laboratory manual*. New York, NY, USA: Cold Spring Harbor Laboratory Press.

Schaefer D, Zyrd J-P, Knight CD, Cove DJ. 1991. Stable transformation of the moss *Physcomitrella patens*. *Molecular & General Genetics* **226**: 418–424.

Schaefer DG, Delacote F, Charlot F, Vrielynck N, Guyon-Debast A, Le Guin S, Neuhaus JM, Doutriaux MP, Nogué F. 2010. RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss *Physcomitrella patens*. *DNA Repair* **9**: 526–533.

Schaefer DG, Zryd J-P. 1997. Efficient gene targeting in the moss *Physcomitrella patens*. *Plant Journal* **11**: 1195–1206.

Schulte J, Erxleben A, Schween G, Reski R. 2006. High throughput metabolic screen of Physcomitrella transformants high throughput metabolic screen of Physcomitrella transformants. *The Bryologist* **109**: 247–256.

Schween G, Egener T, Fritzowsky D, Granado J, Guitton M, Hartmann N, Hohe A, Holtorf H. 2005. Large-scale analysis of 73 329 *Physcomitrella* plants transformed with different gene disruption libraries : production parameters and mutant phenotypes. *Plant Biology* **7**: 228–237.

St Johnston D. 2002. The art and design of genetic screens: *Drosophila melanogaster*. *Nature Reviews Genetics* **3**: 176–188.

Stevenson SR, Kamisugi Y, Trinh CH, Schmutz J, Jenkins JW, Grimwood J, Muchero W, Tuskan GA, Rensing SA, Lang D *et al.* 2016. Genetic analysis of *Physcomitrella patens* identifies *ABSCISIC ACID NON-RESPONSIVE*, a regulator of ABA responses unique to basal land plants and required for desiccation tolerance. *Plant Cell* **28**: 1310–1327.

Sundaresan V. 1996. Horizontal spread of transposon mutagenesis: new uses for old elements. *Trends in Plant Science* **1**: 184–190.

Tadege M, Wen J, He J, Tu H, Kwak Y, Eschstruth A, Cayrel A, Endre G, Zhao PX, Chabaud M *et al.* 2008. Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume *Medicago truncatula*. *Plant Journal* **54**: 335–347.

Trouiller B, Charlot F, Choinard S, Schaefer DG, Nogué F. 2007. Comparison of gene targeting efficiencies in two mosses suggests that it is a conserved feature of Bryophyte transformation. *Biotechnology Letters* **29**: 1591–1598.

Trouiller B, Schaefer DG, Charlot F, Nogué F. 2006. MSH2 is essential for the preservation of genome integrity and prevents homeologous recombination in the moss *Physcomitrella patens*. *Nucleic Acids Research* **34**: 232–242.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research* **40**: e115.

Urbański DF, Matolepszy A, Stougaard J, Andersen SU. 2012. Genome-wide *LORE1* retrotransposon mutagenesis and high-throughput insertion detection in *Lotus japonicus*. *Plant Journal* **69**: 731–741.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Sequences of the 22 cloned SSAP bands for Tnk23 and four for mini-Tnt1 corresponding to the 5′ Tnt1 insertion sites.

Fig. S2 Cloning strategy to obtain the plasmids used for the mini-Tnt1 two-component transposition system.

Fig. S3 Transcriptional profile of genes disrupted by or close to Tnt1 and mini-Tnt1 insertions.

Table S1 List of primers used in this study

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Agraïments

En primer lloc, voldria agrair al Pep per la seva confiança que ha tingut en mi des del primer moment i haver apostat per aquesta tesis. Moltes gràcies per la teva ajuda, paciència i per haver-me ensenyat a tenir esperit crític en la ciència.

Moltes gràcies per tot Pep!

A tots els companys de grup, tant els que ja no hi són com els que es queden, ha estat un plaer poder coincidir i treballar plegats! Moltes gràcies Jordi, Elizabeth, Pol, Bea, Marc, Roger, Fabio, Pedro i Ankita.

A los compañeros del lab 1.02, gracias por vuestras risas, locuras y compañerismo, ha sido un placer compartir con vosotros el lab! Muchas gracias Luis, Andreita, Minky, Soraya y Paula!

Como no, agradecer a los "Marujos y Marujas del CRAG", muchas gracias a todos por los momentos vividos dentro y fuera del CRAG. A las Marujas, des de la más pequeña a la superior gracias por todos esos desayunos y cotilleos :)

Agrair a la Chupipandi, moltes gràcies per la vostra amistat de tants anys i que per molts anys duri! Moltes gràcies Rebeca per aquesta portada i contraportada!

Per últim, voldria agrair a la meva família pel seu suport incondicional i la seva paciència que en molts moments s'ha agraït.

Moltes gràcies pare i mare pel vostre suport i haver-me ajudat arribar fins aquí!